PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ^{6} :
C12N 9/58, 15/55, C07K 16/14

A1

(11) International Publication Number: WO 98/39424

(43) International Publication Date: 11 September 1998 (11.09.98)

(21) International Application Number: PCT/GB98/00704

(22) International Filing Date: 5 March 1998 (05.03.98)

(30) Priority Data:
9704559.5 5 March 1997 (05.03.97) GB

(71) Applicant *(for all designated States except US)*: ISIS INNOVATION LIMITED [GB/GB]; 2 South Parks Road, Oxford OX1 3UB (GB).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: WAKEFIELD, Ann, Elizabeth [GB/GB]; Park View, Woodstock Road, Charlbury, Oxford OX7 3ET (GB).

(74) Agent: PRIVETT, Kathryn, L.; Stevens, Hewlett & Perkins, 1 Serjeants' Inn, Fleet Street, London EC4Y 1LL (GB).

(81) Designated States: JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: DNA ENCODING *PNEUMOCYSTIS CARINII* PROTEASE

(57) Abstract

The invention relates to a novel *Pneumocystis carinii* protease with counterparts in *P.carinii* infecting various different species, including human, as well as nucleic acids encoding it.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

DNA ENCODING *PNEUMOCYSTIS CARINII* PROTEASE

This invention relates to a novel *Pneumocystis carinii* protease and to nucleic acids encoding it. The invention also relates to vectors containing the nucleic acids, to cells transformed with the vectors and to antibodies specific for the protease. In addition, the invention describes uses of all of the above.

The fungal pathogen *Pneumocystis carinii* causes potentially fatal pneumonia in the immunocompromised, including those receiving immunosuppressive therapy for organ transplantation, those with advanced malignancy and in particular those with HIV infection. The lack of an effective *in vitro* culture system still remains a major obstacle in the understanding of the biology of *P.carinii* and its interactions with its host. Molecular techniques have been employed in the study of the organism, and a number of genes have now been cloned. Among these is the multi-gene family encoding the major surface glycoprotein, (MSG or gpA) of the parasite.

The *P.carinii* major surface glycoprotein is highly mannosylated and is antigenically distinct in organisms isolated from different mammalian host species (Lundgren *et al.*, 1991; Gigliotti, 1992). The *MSG* multi-gene family has been identified in the genome of *P.carinii* sp. f. *carinii* (rat-derived *P.carinii*) Kovacs *et al.*, 1993; Wada *et al.*, 1993; Sunkin *et al.*, 1994), *P.carinii* sp. f. *mustelae* (ferret-derived *P.carinii*) (Haidaris *et al.*, 1992; Wright *et al.*, 1995), *P.carinii* sp. f. *hominis* (human-derived *P.carinii*) (Stringer *et al.*, 1993) Garbe & Stringer, 1994) and *P.carinii* sp. f. muris (mouse-derived *P.carinii*) (Wright *et al.*, 1994). The different copies of *P.carinii* sp. f. *carinii MSG* genes are of similar size but heterogeneous in sequence. They have been found on multiple chromosomes and often organised in tandem arrays. The majority of *MSG* genes are located in the subtelomeric regions of the *P.carinii* sp. f. *carinii* chromosomes (Underwood *et al.*, 1996; Sunkin & Stringer, 1996). The expression of *MSG* genes has been shown to be mediated by the upstream conserved sequence (UCS) which is found on a single chromosome situated in the subtelomeric region. Different copies of *MSG* have been shown to be linked to the UCS. It has been postulated that this differential expression of *MSG* may occur in a strategy to evade the immune response of the host by antigenic variation (Wada *et al.*, 1995; Sunkin & Stringer, 1996).

Presently there are two standard treatments for *Pneumocystis* pneumonia, namely pentamidine or cotrimoxazole. These drugs were originally used because it was thought that *Pneumocystis* was a protozoan; only recently has genetic sequence analysis placed it in the fungal kingdom. Despite its classification as a fungus, *Pneumocystis* does not respond to the usual anti-fungal drugs and hence the drug regimes have remained all but unchanged. These regimes are particularly unpleasant with many patients reacting adversely, thus requiring a switch in treatment. Thus AIDS patients in particular would benefit from the development of new anti-*Pneumocystis* therapies since a high proportion of AIDS patients suffer adverse side effects, and many have multiple episodes of *P. carinii* pneumonia due to their decreasing CD4+ lymphocyte count and persistence of immune suppression.

Recently, a novel family genes from *P.carinii* sp. f. *carinii* has been described (Lugli and Wakefield 1996). The genes are found in the subtelomeric regions of the *P.carinii* sp. f. *carinii* genome, and show homology to protease genes from a number of fungi.

Wada and Nakumura (1994) describes the discovery of an open reading frame (designated ORF-3) encoding a protein of unknown function in *P.carinii* sp. f. *carinii* and located close to the *MSG* genes. The sequence given (DDBJ/EMBL/GenBank accession no. D31909 and

D17441) corresponds to a portion of the genes discussed above (Lugli and Wakefield 1996).

It has now been discovered that there is a *P.carinii* sp. f. *hominis* counterpart to the family of genes in the rat-derived *P. carinii* species referred to above, the human-derived *P. carinii* species having at least 50% difference to the rat-derived *P. carinii* species in its nucleotide sequence. The novel multi-gene family is known as *PRT1* (Protease 1); the genes show high levels of homology with the subtilisin-like serine proteases.

The subtilisin-like serine proteases are a group of endoproteases which have been characterised from a wide variety of organisms including bacteria, fungi and higher eukaryotes. They have been found to function in the specific endoproteolytic processing of pro-proteins at cleavage sites of paired basic amino acid residues, to generate regulatory proteins in a mature and biologically active form. The pro-hormone processing enzyme kexin, encoded by the *KEX2* gene of *Saccharomyces cerevisiae* has been characterised and found to cleave the precursors of the α-mating factor and the killer toxin (Fuller *et al.*, 1989). Genes encoding a similar processing endoprotease have been identified in a number of other fungi, the *KEX1* gene from the yeast *Kluyveromyces lactis* (Tanguy-Rougeau *et al.*, 1988), the gene encoding the *KEX2*-related protease (krp) from *Schizosaccharomyces pombe* (Davey *et al.*, 1994) and the *XPR6* gene from *Yarrowia lipolytica* (Enderlin & Ogrydziak, 1994). Mammalian homologues have also been identified incluidng the human *fur* gene (fes upstream region) in the region upstream of the fes proto-oncogene, encoding the enzyme furin (van den Ouweland *et al.*, 1990). The genes Dfur1 and Dfur2 from the insect *Drosophila melanogaster* encoding furin-like proteins (Roebroek *et al.*, 1992) and the *bli-4* gene from the nematode *Caenorhabditis elegans* have also been studied. Other members of the subtilisin-like serine protease family have been identified and the specific endoproteolytic activity of some of them has been elucidated. However for many others, the precise biological function has not yet been determined.

The *PRT1* gene product may be a specific endoproteolytic processing enzyme, such as is seen in other subtilisin-like serine proteases. Given that in genetic organisation some copies of *PRT1* are generally found in the subtelomeric region, just downstream from the *MSG* gene, the PRT1 protein encoded by these genes may be involved in the processing of MSG to its mature form. The multicopy nature of the *PRT1* gene may reflect the need for processing of enzymes of different specificity for the different types of MSG. Whatever its precise role, the activity of the PRT1 protein is undoubtedly essential to the viability and therefore the pathogenesis of *P.carinii*.

Recently, there has been considerable interest in targeting proteases, for the control of a number of different diseases and in particular HIV infection. Combination therapies for HIV treatment employ protease inhibitors; a large variety of protease inhibitors are therefore available for testing against new proteases.

The Invention

Part of the catalytic domain of a *PRT1* gene has been cloned, sequenced and characterised from three types of the host specific fungal pathogen *P.carinii*, namely *P.carinii* sp. f. *rattus* (rat variant), *P.carinii* sp. f. *muris* (mouse) and *P.carinii* sp. f. *hominis* (human). The newly discovered human-infecting *P.carinii PRT1* catalytic domain sequence is shown in figure 1 and nucleotide sequence alignments for rat *P. carinii*, rat variant *P. carinii*, mouse *P. carinii* and human-infecting *P.carinii PRT1* clones are shown in figure 2. These will enable the sequencing of the remaining parts of a *PRT1*, using techniques known to those skilled in the art of molecular biology.

The invention therefore provides in one aspect an isolated DNA comprising part or all of a *PRT1* gene of a non-rat infecting species of *Pneumocystis carinii.*

The invention also provides an isolated DNA comprising a sequence shown in figure 1, or a non-rat *P. carinii* sequence shown in figure 2, or a sequence which hybridises to either of these under stringent conditions.

In further aspects, the invention provides recombinant vectors containing *PRT1* DNA sequences as described herein, and recombinant polypeptides which are part or all of a *PRT1* gene product, encoded by the vectors.

In another aspect, the invention provides synthetic peptides corresponding to antigenic portions of a *PRT1* gene product.

In further aspects, the invention provides a method of producing antibodies specifically immunoreactive with a *P.carinii* protease, which method comprises using a recombinant polypeptide or a synthetic peptide as described herein to generate an immune response; and antibodies produced by the method.

In another aspect, the invention provides a method of screening for anti-*Pneumocystis carinii* compounds, which method comprises providing a source of a recombinant polypeptide expressed by part or all of a *PRT1* gene or cDNA, and contacting the compound with the recombinant polypeptide.

In another aspect, the invention provides an engineered cell transfected with a recombinant vector containing *PRT1* DNA sequences as described herein.

In another aspect, the invention provides an engineered cell line expressing a recombinant polypeptide from part or all of a *PRT1* gene or cDNA, useful in a method of screening for anti-*P.carinii* compounds such as protease inhibitors effective against *P.carinii.*

In another aspect, the invention provides a *P.carinii* protease isolated using an antibody specifically immunoreactive with a *P.carinii* protease, as described herein.

In another aspect, the invention provides *PRT1* clones for part or all of a human-infecting *P.carinii PRT1* gene from the *PRT1* multi-gene family.

A part of the PRT1 gene as referred to herein may be for example a fragment of the gene which codes for a specific domain such as the catalytic domain, or it may be a shorter sequence such as a sequence not less than 15 nucleotides in length or not less than 20 nucleotides in length. Sequences of about 15 or about 20 nucleotides in length are generally the shortest practical length of oligonucleotide useful as a sequence specific primer or probe. That is, these are generally the shortest lengths of sequence that will hybridise specifically to a gene sequence under stringent conditions.

Within the *PRT1* multi-gene family will be related genes which will be easily identifiable as such by those skilled in the art, but which may nevertheless differ in location, function and sequence. It will be evident that all members of the *PRT1* multi-gene family, which members may variously be described as different genes in the family or as different copies of the *PRT1* gene, are included within the scope of the invention.

Known methods to mutate or modify nucleic acid sequences can be used in conjunction with this invention to generate useful *PRT1* mutant sequences. Such methods include but are not limited to point mutations, site directed mutagenesis, deletion mutations, insertion mutations, mutations obtainable from homologous recombination, and mutations obtainable from chemical or radiation treatment.

Furthermore, recombinant DNA techniques are available to mutate the DNA sequences described herein, to link these DNA sequences to expression vectors and express the PRT1 protein or part of the protein eg. the catalytic domain or the P-domain.

In the attached figures:

Figure 1 shows the genomic DNA sequence of part of the catalytic domain of *PRT1* from *P.carinii* sp. f. *hominis*. (SEQ ID NO: 22)

Figure 2 shows DNA sequence alignments for part of the catalytic domain of *PRT1* from *P.carinii*. (Found in GenBank AF001305, GenBank AF001304, and SEQ ID NOS: 23 – 29, in the order in which they appear).

Figure 3 shows amino acid sequence alignments of part of the catalytic domain of *PRT1*, translated from the nucleotide sequences in figure 2. (Found in GenBank and SEQ ID NOS: as for Figure 2).

Figure 4 shows alignment of *P.carinii PRT1* derived amino acid sequences from *P.carinii* sp. f. *carinii* clones. (Found in GenBank AF001305, GenBank AF001304 and SEQ ID NOS: 30, 31, 33 – 47, 32, 48 – 50).

Figure 5 shows DNA sequence alignments for *P.carinii* sp.f. *carinii PRT1* clones. (Found in GenBank AF001305, GenBank AF001304 and SEQ ID NOS: 30 – 32)

Figure 6 shows a schematic representation of the *P.carinii* sp. f. *carinii PRT1* gene.

Figure 7 shows expressed recombinant PRT1 fragments.

By analogy to *P.carinii* sp. f. *carinii* there are expected to be many copies of the *PRT1* gene within the *P.carinii* sp. f. *hominis* genome. Some of these copies may be significantly different and form a number of different sub-types. They will all, however, be classed as members of the *PRT1* multi-gene family by virtue of homology at some domains of the gene, for example the catalytic domain.

Seven different domains have been identified to date in the *P.carinii* sp. f. *carinii PRT1* amino acid sequence, namely:

i) N-terminal hydrophobic domain

ii) Pro-domain

iii) Catalytic domain

iv) P-domain

v) Proline-rich domain

vi) Serine-threonine rich domain

vii) C-terminal hydrophobic domain

The *P.carinii* sp. f. *hominis* homologues may have fewer, the same number or more domains. Although some domains in some members of *P.carinii* sp. f. *hominis PRT1* gene family may be absent or some extra domains may be present, these genes will still be considered to be members of the *PRT1* multi-gene family.

The proteins encoded by different copies of this gene family may have a variety of different functions, including:

i) as a constituent of the outer cell surface of the parasite, and attached to the cell membrane by a glycosyl-phosphatidylinositol (GPI) anchor

ii) the proteolytic processing within a *P. carinii* sub-cellular organelle of the *P.carinii* major surface glycoprotein (MSG) to its mature form, possibly at a conserved dibasic amino acid site in the upstream conserved sequence of MSG

iii) in the interaction of the parasite with its host, forming a specific ligand on the parasite cell surface which binds to a host receptor molecule

There may be other functions of the members of this gene family which have not yet been recognised. These may include functioning as a protease on as yet unidentified pro-proteins, or as a structural glycoprotein at some life-cycle stage of the parasite.

It has been demonstrated that the protease is a surface protease.

Therapeutic intervention

The PRT1 protein presents a target for a variety of different therapeutic interventions, which may include:

i) <u>Inhibitors of protease activity</u>

It is postulated that the proteolytic activity of PRT1 is essential for the viability of the parasite. The predicted structure of the catalytic domain of the PRT1 protein suggests that there are subtle differences compared to other such proteases so far studied. These differences may be exploited in the design of specific drugs, with less toxic side-effects than seen in the present available treatments.

ii) <u>Vaccines</u>

Available data indicates that some copies of PRT1 may comprise a major surface antigen and therefore provide a potential target for vaccine development.

iii) <u>Immunotherapy</u>

Passive immunisation with antibodies to PRT1 may be protective.

iv) <u>Analogues</u>

Analogues designed to imitate PRT1 may be active in blocking the adherence of *P.carinii* organisms to a receptor on the human cells.

Identification of a subtilisin-like serine protease in *P.carinii* sp. f. *carinii*

METHODS

P.carinii DNA extraction

P.carinii infection was induced in Sprague Dawley rats by steroid immunosuppression. The organisms were isolated and purified from infected rat lung tissue by the method described by Peters *et al.*, (1992). Genomic *P.carinii* DNA was extracted by digestion with proteinase K (1 mg/ml) in the presence of 0.5% SDS and 10mM EDTA, pH8.0, at 50°C for 16h, followed by phenol:chloroform extraction and ethanol precipitation. *P.carinii* DNA for use in PFGE experiments was prepared in SeaPlaque GTG agarose as described by Banerji *et al.*, (1993).

For oligonucleotide primers, see Table 1 and Lugli *et al* 1997.

Isolation of copies of the *PRT1* gene from *P.carinii* sp. f. *carinii* genomic and cDNA libraries

A copy of the *PRT1* gene was isolated from an unamplified genomic library from *P.carinii* sp. f. *carinii* constructed in λEMBL3 (Banerji *et al.*, 1993). The library was screened with a cDNA clone containing a region of a *P.carinii* sp. f. *carinii MSG* gene (GenBank Accession number GBPLN:PMCANTIA, donated by Dr C J Delves and Dr F Volpe). A relatively high number of recombinant plaques gave positive hybridization signals compared to the positive recombinant plaques when the library was screened with a probe derived from the single copy *arom* locus (Banerji *et al.*, 1993). Five recombinant phages were isolated from the tertiary screen and the DNA was subcloned into the plasmid vector pBluescriptII.

In order to isolate a full cDNA clone, a *P. carinii* sp. f. *carinii* cDNA library constructed in λZAPII (donated by Dr CJ Delves and Dr F Volpe, see Dyer *et al.*, 1992), was screened with PCR products derived from amplification of the 5' end of the gene with oligonucleotide primer pair pcprot9 and prp4r (9/4r product), and of the 3' end of the gene with pcprot13/RI and pcprot12/RI (13/12 product). The primary screening was carried out using both probes, and the secondary and tertiary screens were carried out using only the 9/4r product. The number of positive clones when screening the cDNA library with the two probes appeared to be relatively high when compared to the number obtained using a single copy gene. Four recombinant phage isolated from the cDNA library were partially characterized. The recombinant DNA was recovered from the λ phage by *in vivo* excision as pBlueScript plasmid DNA. The size of the recombinant DNA ranged from 2.7kb to 2.9kb, and sequence analysis revealed that all four clones contained a polyA tail. One recombinant, 73j was selected for further analysis and the recombinant DNA was sequenced in full from both strands.

DNA amplification

Oligonucleotide primers were designed to various regions of the *P.carinii PRTI* nucleotide sequences. Some oligonucleotides had an *Eco*RI restriction endonuclease site incorporated at the 5' end to facilitate cloning of the amplification products into *Eco*RI-digested plasmid vectors pBluescript SK(-) (Stratagene) or pUC18 (Pharmacia). The final concentration of the amplification reaction mix was 50mM KCl, 10mM Tris (pH8.0), 0.1% Triton X-100, 3mM $MgCl_2$, 400µM (each) deoxynucleoside triphosphate, 1µM oligonucleotide primer and 0.025 U Taq polymerase ml^{-1} (Promega, UK). With primer pair pcprot9 and pcprot10, forty cycles of amplification was performed at 94°C for 1.5 min., 53°C for 1.5 min., and 72°C for 2.0 min. With primer pair pcprot9 and pcprot4r the same conditions were used, except an annealing temperature of 50°C was used. With all other primer pairs, ten cycles of amplification were carried out at 94°C for 1.5 min., 55°C for 1.5 min., and 72°C for 2.0 min, followed by 30 cycles of 94°C for 1.5 min., 63°C for 1.5 min., and 72°C for 2.0 min. Negative controls were included in each experiment.

The entire putative gene was amplified as three overlapping fragments, Prp5e (1626 bp), M14 (1279 bp) and Prp2g (251 bp). Oligonucleotide primer pairs pcprot9 with pcprot10, followed by pcprot6/RI with pcprot4/RI were used in a nested PCR to amplify the 5' fragment, designated Prp5e, of length 1626 base pairs (bp). The second portion, called M14, spanning 1279 bp of the central region of *PRTI*, was amplified using a nested PCR with primer pairs pcprot2/RI with pcprotl4/RI, followed by pcprot7/RI with pcprot12/RI. The third fragment, Prp2g, encompassing the 3' end of the sequence (251 bp), was amplified using oligonucleotides primers pcprot13/RI and pcprot14/RI (Table 1 and Lugli *et al* 1997).

Five different overlapping regions of the *PRTI* gene were also amplified, cloned and the DNA sequences were determined. The first region amplified with primer pair pcprot1/RI and pcprot3/RI spanned approximately half of the subtilisin-like catalytic domain, the second region amplified with primer pair pcprot2/RI and pcprot4/RI spanned the end of the subtilisin-like catalytic domain and the start of the P-domain, the third region amplified with primer pair pcprot7/RI and pcprot8/RI spanned the P-domain, the fourth region amplified with primer pair 36ex/RI and Pt3/RI spanned the proline-rich domain and the fifth region amplified with primer pair pcprot13/RI and pcprot 14/RI spanned the C-terminal hydrophobic domain. The sequences Prp1a, Prp3a, Prp7a, Prp2c, Prp3c, Prp4c, Prptaf2, Prpf4, Prp5f, Prpg3 and Prp5g were amplified from the *P. carinii* cDNA library, and sequences Pcr-19, Pcr-14, Pcr-5, Pcr-3, Pcr-1, Lam-1 and Prpg4 from the *P.carinii* genomic DNA (Figure 4).

DNA sequence analysis

DNA sequence analysis was performed using the dideoxy chain termination method. Sequence data was obtained in full from both strands for all sequences. Analysis of the sequence data was carried out using the University of Wisconsin Genetics Computing Group (UWGCG) Sequence Analysis Software Package, Version 8, 1994, Genetics Computer Group, Madison, Wisconsin.

Pulsed Field Gel Electrophoresis

P. carinii sp. f. *carinii* organisms were isolated from an infected rat lung and the chromosomes were separated by pulsed field gel electrophoresis (PFGE), using a Contour Clamped Homogeneous Electric Field (CHEF) DRII apparatus (Bio-Rad, UK) operated at 4^{O}C. Electrophoretic separation was achieved using 0.9% Seakem agarose gel with initial switching time of 10 sec increasing to a final switching time of 60 sec at 180 V for 48 hours. A karyotype corresponding to *P.carinii* sp. f. *carinii* form 1 was observed (Cushion *et al.*, 1993).

Southern hybridisation

Southern blotting and hybridization were carried out using standard techniques (Sambrook *et al.*, 1989). PFGE blots were hybridised with three probes derived from different domains of the *PRT1* gene. The product 9/4r was derived from amplification of the 5' end of the *PRT1* gene with primer pair pcprot9 and pcprot4r/RI, product 2/4 from amplification of the central catalytic region with primer pair pcprot2/RI and pcprot4/RI, and product 13/12 from amplification of the 3' end of the gene with primer pair pcprot13/RI and pcprot12/RI. The amplification products were gel-purified (GeneCleanII, BIO101) and labelled with [α-^{32}P]-dCTP by random priming (Megaprime, Amersham). Hybridisation was carried out at 45°C and stringency washing at 60°C in 0.2xSSC and 0.1% SDS.

Southern blots of genomic *P.carinii* DNA digested with restriction endonuclease *Pst*I or *Bam*HI were probed with oligonucleotide probes pcprot3/RI, pcprot5/RI, pctel2, and msgterm, labelled with [γ-^{32}P]-dATP using polynucleotide kinase. Hybridisation was carried out at 46°C and stringency washing at 52°C in 5xSSC and 0.5% SDS.

RESULTS

Analysis of DNA and deduced amino acid sequence of copies of the *PRT1* gene

We have identified a family of genes in the *P.carinii* sp. f. *carinii* genome which shows homology to the subtilisin-like serine proteases. We have named this gene family *PRT1* (protease 1). A copy of the *PRT1* gene (Paga) was isolated from a *P.carinii* genomic library, the open reading frame (3069bp) containing seven short putative intervening sequences. A copy of the *PRT1* gene (73j) was also isolated from a cDNA library, of length 2370bp. Portions of the gene were amplified by PCR from the cDNA library as three overlapping fragments, at the 5' end (Prp5e), the central region (M14) and the 3' end (Prp2g). Five other regions of the gene were also amplified, from either the *P.carinii* cDNA or genomic libraries.

Analysis of the DNA sequence of the copy of the *PRT1* gene from the genomic library, *PRT1*(Paga), and of the copy from the cDNA library, *PRT1*(73j), confirmed the presence of seven short introns in the genomic DNA sequence. The introns ranged in length from 38 bp to 45 bp, with a base composition ranging from 71% to 84% A+T. In all seven introns, the dinucleotide GT was present at the 5' splice donor site and AG at the 3' splice acceptor site. The sequence YTRAT, which has been identified as the putative lariat forming motif in other *P.carinii* sp. f. carinii introns (Zhang & Stringer, 1993), was present in the first, second, fourth, fifth and seventh intron. The eukaryotic lariat consensus sequence, YYRAY, was identified in the third and sixth intron.

The sequence of the cDNA clone, *PRT1*(73j), contained an open reading frame of 2370bp, which on translation resulted in a peptide of 790 amino acids (Figure 4). The deduced amino acid sequence was compared to sequences in the GenBank and EMBL databases and showed homology to fungal and other eukaryotic subtilisin-like serine proteases. The A+T content of the ORF was 64%, with a high A+T content at the third base position of the codons. The base composition of the 5' upstream sequence was 74% A+T, and the 3' downstream sequence was 75% A+T. A consensus polyadenylation signal, AATAAA, was observed 68bp downstream of the stop codon.

The deduced amino acid sequence of the genomic clone *PRT1*(Paga), the cDNA clone *PRT1*(73j), the three fragments obtained by PCR amplification of the cDNA library and the other recombinant clones generated by DNA amplification were compared (Figure 4). Several regions of homology were found and also a number of regions in which significant divergence was observed. These data suggested that the sequences were derived from different copies of the *PRT1* gene.

Comparison with other subtilisin-like serine proteases

The deduced amino acid sequence of the cDNA clone *PRT1*(73j) was aligned with nine other subtilisin-like serine proteases including fungal, mammalian, insect and nematode sequences. The PRT1 sequences showed homology with all the other sequences, with a high level of homology in the subtilisin-like catalytic domain. The three essential residues of the catalytic active site, aspartic acid (Asp_{214}), histidine (His_{252}) and serine (Ser_{423}) were conserved in all the PRT1 sequences. The highest levels of homology between all the sequences were around these residues.

The structural organisation of the fungal sequences showed domains characteristic of this class of processing endoproteases, a hydrophobic signal sequence, a pro domain that may be cleaved by autoproteolysis, a subtilisin-like catalytic domain, a P-domain which is known as such because it is essential for proteolytic activity, a serine/threonine-rich domain which may potentially be modified by O-linked glycosylation, a carboxy-terminal hydrophobic trans-membrane domain and a C-terminal tail with acidic residues (Van de Ven *et al.*, 1993) The *P.carinii* PRT1 sequences showed a putative similar structural organisation but unlike the nine other subtilisin-like serine proteases, they also had a proline-rich domain preceeding the serine-threonine rich domain and the C-terminal hydrophobic domain (Figure 6). The *P.carinii* PRT1(73j) sequence had a hydrophobic signal sequence at the N-terminus, followed by a putative pro-domain, a subtilisin-like catalytic domain from Ser_{171} to His_{474}, a P-domain from residue Tyr_{475} to Ser_{631}, a proline-rich domain from residue Pro_{641} to Pro_{707}, a serine-threonine rich domain from residues Thr_{708} to Ser_{765}, and a carboxy-terminal hydrophobic domain from residues His_{771} to Phe_{790}.

Analysis of subtilisin-like catalytic domain

The three-dimensional structures of four subtilisin-like serine proteases have been determined, subtilisin BPN'/Novo from Bacillus amyloliquefaciens (Hirono *et al.*, 1984; Bott *et al.*, 1988), subtilisin Carlsberg from *B. licheniformis* (McPhalen & James, 1988), thermitase from *Thermoactinomyces vulgaris* (Gros *et al.*, 1989; Teplyakov *et al.*, 1990) and proteinase K from *Titirachium album* (Betzel *et al.*, 1988). The amino acid sequence of these four proteases has been compared to that of 31 other subtilisin-like serine proteases isolated from bacteria, fungi and higher eukaryotes and the essential core structure of the catalytic domain of this group of molecules has been identified (Siezen *et al.*, 1991).

We have compared the deduced amino acid sequence of the *P.carinii PRT1*(73j) gene with the multiple sequence alignment of the other subtilisin-like serine proteases and have identified the three essential residues of the catalytic active site aspartic acid, histidine and serine in the PRT1 sequence (Asp_{214}, His_{252} and Ser_{423}). On the basis of the sequence alignment, the *P.carinii* PRT1 sequence could be assigned to the class 1 subtilases, within the subgroup I-E which contained the pro-hormone processing proteases from yeasts and higher eukaryotes (Siezen *et al.*, 1991).

Eight α-helical domains and nine β-sheet regions have been defined as the structurally conserved regions within the essential core structure. The variable regions which connect the core segments have been found to differ both in length and in amino acid sequence (Siezen *et al.*, 1991). High levels of homology were observed between the PRT1 sequences and the other sequences in the regions of the two conserved internal helices, helix C (residues 252 to 262) and helix F (residues 422 to 438). Eleven amino acid residues have previously been found to be totally conserved in all the characterized subtilisin-like serine proteases, and most but not all are conserved in the PRT1 sequences. These amino acid residues are at the active site, Asp_{214}, His_{252} and Ser_{423}, [found in all the PRT1 sequences except PRT1(Prp7a)] and in the internal helices at residues Gly_{253}, Gly_{258}, Pro_{427}. The residues Ser_{310}, Gly_{312}, Gly_{351}, Gly_{421} and Thr_{422}, involved in substrate binding, were conserved in all the PRT1 sequences, except Thr_{422} which was found only in two sequences generated by PCR, PRT1(Prpla) and PRT1(Prp7a).

In addition to the totally conserved residues, seven other amino acid residues have been identified which are highly conserved, of these six were conserved in the *P.carinii* PRT1 sequences and included the oxyanion hole residue (Asn_{352}), residues near the active site, Gly_{216}, Thr_{254}, and also residues Gly_{205}, Gly_{271} and Gly_{343}. Seven conserved cysteine residues were found in all the *P.carinii* PRT1 sequences, Cys_{256}, Cys_{288}, Cys_{309}, Cys_{359}, Cys_{389}, Cys_{391} and Cys_{415}. Nineteen variable regions, generally located in loops on the surface of the molecule, have been identified in the subtilase family, of which 14 were found in the *P.carinii* PRT1 sequences. Three positions have been identified at which charge is totally conserved in all the subtilisin-like proteases examined, and these were also conserved in the *P.carinii* PRT1 sequences, the positive charge on Arg_{282} and the negative charges on residue Asp_{214} (active site) and Asp_{223}.

It has been proposed that the high specificity of the class I-E subtilisin-like serine proteases for paired basic residues Lys-Arg or Arg-Arg may be facilitated by a high density of negative charge at the substrate-binding face, provided by nine highly conserved Asp residues and one Glu residue (Siezen *et al.*, 1991). Two of the Asp residues, Asp_{353} and Asp_{409} were found in all the *P.carinii* PRT1 sequences and also the Glu_{293}. In addition, four other Asp residues were found in some but not all of the copies of PRT1.

Analysis of the domains flanking the subtilisin-like catalytic domain

The putative domains of the PRT1(73j) polypeptide are summarised in Figure 6. A hydrophobicity plot of the PRT1(73j) sequence revealed a hydrophobic region at the N-terminus suggesting that this may be a signal sequence. Residues 1 to 23 of the N-terminus of the sequence showed a high level of homology to the N-terminus of the *P.carinii* sp.f. *carinii* multifunctional folic acid synthesis *fas* gene which encodes dihydroneopterin aldolase, hydroxymethyldihydropterin pyrophosphokinase and dihydropteroate synthase (Volpe *et al.*, 1992, 1993). This region was followed by the presumptive pro-domain, which may be cleaved by autocatalysis. Potential autocatalytic sites of paired basic residues were identified in the PRT1(Paga) and PRT1(Prp5e) sequences at Lys_{115} - Arg_{116} and Arg_{136} - Arg_{137}, but were absent in the PRT1(73j) sequence. Five other semi-conserved autocatalytic sites were found in some copies, but not all, of the *P.carinii* PRT1 sequences, two in the catalytic domain (Lys_{400} - Arg_{401}, Arg_{473} - Arg_{474}), three in the P-domain (Arg_{521} - Arg_{522}, Arg_{555} or Lys_{555} - Arg_{556}, Arg_{576} - Arg_{577}). One potential autocayalytic site at the start of the carboxy-terminal hydrophobic region (Lys_{769} - Arg_{770}), which was found in all the sequences. The PRT1(73j) sequence contained two of the potential autocatalytic sites, Arg_{576} - Arg_{577} and Lys_{769} - Arg_{770}.

The PRT1 sequences showed homology with the other subtilisin-like serine proteases in the region of the P-domain, the highest homology being with the derived amino acid sequence of the *S. pombe krp* gene. Four potential sites for N-linked glycosylation were observed in all the PRT1 sequences, three in the subtilisin-like catalytic domain (Asn_{194}, Asn_{277}, Asn_{442}), and one in the P-domain (Asn_{603}).

A serine-threonine rich region was also identified in the PRT1(73j) sequence from residue Thr_{708} to Ser_{765}, and the hydrophobicity plot of the PRT1(73j) sequence revealed a hydrophobic region at the C-terminal end, residues His$_{771}$ to Phe$_{790}$, suggesting a membrane-associated domain. Unlike most other serine protease sequences, however, all the copies of the PRT1 polypeptide contained a proline-rich region downstream of the P-domain.

Genetic organization of the PRT1 multi-gene family

Analysis of the alignments of the DNA and the deduced amino acid sequences of copies of the *PRT1* gene from genomic DNA, the cDNA sequence and the three fragments obtained by PCR of the cDNA library revealed domains in the *PRT1* gene which were highly conserved and also regions where significant divergence was observed, again suggesting that *PRT1* comprises a multi-gene family (Figure 4). The subtilisin-like catalytic domain and the P-domain appeared to be conserved whereas high levels of heterogeneity were observed in the proline-rich domain and the C-terminal domain. The variation in this region was both in length and in sequence. A number of repeated DNA sequence motifs were found in the proline-rich region. Nucleotide sequences encoding polyproline were found in all the sequences, and also the dipeptides Pro-Glu and Pro-Gln and the tetrapeptides Pro-Glu-Pro-Gln and Pro-Glu-Thr-Gln. The order and number of tandem repeats varied in each sequence. The overall length of this region varied from approximately 67 amino acid residues in the shortest sequence, PRT1(73j), to 233 residues in the longest sequence, PRT1(M14).

In order to further substantiate the presence within the *P.carinii* genome of multiple copies of the *PRT1* gene, *P.carinii* sp. f. *carinii* chromosomes, separated by pulsed field gel electrophoresis, were analysed by hybridisation with three probes derived from different domains of PRT1. All three probes showed similar patterns of hybridization, anealing at high stringency to all the chromosome bands except for one, the third smallest in size, approximatey 350Kbp. This provided further evidence that the *P.carinii* sp. f. *carinii* genome contained many copies of the *PRT1* gene, which were present on most of the *P.carinii* sp. f. *carinii* chromosomes.

The sequences of the *PRT1* gene family showed high levels of homology with ORF3, which has been demonstrated to be contiguous with a copy of the gene encoding the major surface glycoprotein *MSG100* (Wada & Nakamura, 1994). This gene arrangement was reported in 15 other λ clones, in which a gene showing high homology to ORF3 was located downstream of a copy of *MSG* (Wada & Nakamura, 1994). Most copies of the *MSG* genes have been demonstrated to be located in the *P.carinii* sp. f. *carinii* subtelomeric regions (Underwood *et al.*, 1996; Sunkin & Stringer, 1996). The copy of the *PRT1* gene encoded by the PRT1(Paga) sequence was cloned from a λ EMBL3 genomic library as a single 14kb fragment and was approximately 1150bp downstream of a copy of *MSG*. Four other λ clones isolated from the same library contained a copy of *PRT1* contiguous with a copy of *MSG*.

P.carinii sp. f. *carinii* genomic DNA was digested with either restriction endonuclease *PstI* or *Bam*HI and probed sequentially with four oligonucleotide probes, derived from the 5′ end of *PRT1* gene (pcprot5/RI), from the catalytic domain of the gene (pcprot3/RI), an *MSG* probe (msgterm) and a subtelomeric probe (Pctel2). All probes hybridised to multiple bands. The hybridisation pattern of some of the bands, ranging in size from 7kb to greater than 12kb, were the same for all four probes. However, hybridisation to other fragments was not coincident, with the *PRT1* probes alone hybridising to some high molecular weight fragments and also low molecular weight fragments of less than 7kb.

DISCUSSION

We describe the cloning and characterisation of copies of the *PRT1* multi-gene family from *P.carinii* sp. f. *carinii*. A copy of the *PRT1* gene was isolated from a *P.carinii* sp. f. *carinii* genomic library. A different copy was isolated from a cDNA library, indicating that this copy of the gene was transcribed, and also identifying the presence of seven short introns in the genomic sequence. Consistent with many other *P.carinii* genes, the coding region and the flanking sequences of the *PRT1* sequences showed a strong bias for adenine or thymine, and in particular at the third base position of the codons. Similarly, the presence of short A+T rich introns has been reported in other *P.carinii* genes. In the *PRT1* sequences, the introns were not distributed throughout the gene, but six of the seven introns were found in the subtilisin-like catalytic domain, and the seventh in the P-domain. The introns may play a role in restricting the variation in this region of the gene, whereas no introns were observed in the highly heterogeneous proline-rich region (Rogers, 1985).

The high level of homology of the *P.carinii PRT1* sequences to the subtilisin-like serine proteases, and in particular in the region of the catalytic domain, strongly suggested that this gene encoded a protease of this type. The predicted *P.carinii PRT1* polypeptide sequences possessed the three essential residues of the catalytic active site as well as many other highly conserved motifs. The domain organisation of the *PRT1* gene strongly resembled that of the fungal prohormone processing proteases, with the exception of the proline-rich domain. This proline-rich region is very uncommon in the subtilisin-like serine protease superfamily, although the *KRP6* gene from *Y. lipolytica* is reported to contain a short region of a tetrapeptide repeat, the consensus sequence of the four amino acids being Glu (Asp/Glu) Lys Pro (Enderlin and Ogrydziak, 1994). A proline-rich region has also been found in the carboxy-terminal tail domain of the mammalian serine protease acrosin, a proteolytic enzyme of sperm cells, located in the acrosome at the apical end of the spermatozoan (Klemm *et al.*, 1991).

In the African trypanosome, *Trypanosoma brucei*, a proline-rich domain has been identified in the procyclic acidic repetitive proteins

(PARPs). These proteins are found on the cell surface of the insect form of the parasite and are encoded by a family of polymorphic genes which contain a variable region with heterogeneity both in length and sequence. The variable region contains the proline-rich domain and is primarily composed of the dipeptide Glu-Pro (Roditi *et al.*, 1989).

Unlike any of the other fungal prohormone processing proteases, which appear to be single copy genes, the data reported in this study suggest that the *PRT1* sequence is present in many copies, which are similar but not identical, in the genome of *P.carinii* sp. f. *carinii*. The relatively large number of recombinants present in both the genomic and the cDNA libraries suggested a multi-copy gene and this was substantiated by PFGE data, revealing that at least one copy of a *PRT1* gene was present on all but one of the *P.carinii* chromosomes. Southern hybridisation of restriction endonucleolytic digests of *P.carinii* sp. f. *carinii* DNA probed with *PRT1* sequences also confirmed the presence of many copies of the gene. Analysis of sequence data generated by the amplification of the locus showed heterogeneity, suggesting that a variety of different copies of the gene were present in the *P.carinii* genome. Some domains, including the subtilisin-like catalytic domain and the P-domain, were highly conserved between gene copies, whereas the highest levels of divergence were observed in the proline-rich domain, which varied both in length and in sequence.

Of five genomic clones analyzed in this study, all possessed a copy of *PRT1* contiguous with a *MSG* gene. It has been reported that 15 independent genomic clones which encoded *MSG* were contiguous with the ORF3 sequence, which from our analysis, appears to encode the proline-rich domain of *PRT1* (Wada & Nakamura, 1994). It has been demonstrated that most copies of *MSG* are subtelomeric (Underwood *et al.*, 1996, Sunkin & Stringer, 1996). It is therefore highly likely that many copies of the PRT1 multi-gene family are located in the subtelomeric regions of the *P.carinii* sp. f. *carinii* genome. However PFGE analysis has shown that not every *P.carinii* sp. f. *carinii* chromosome contained a copy of *PRT1*, and the preliminary characterisation of a clone of one of the subtelomeric regions of *P.carinii* sp. f. *carinii* has not revealed a copy of *PRT1* (Underwood & Wakefield, unpublished results). Hybridisation of *MSG* and subtelomeric probes to endonuclease digested *P.carinii* sp. f. *carinii* DNA resulted in positive hybridisation to fragments greater than approximately 7 kb in size. Probes derived from the *PRT1* sequence hybridised to these bands but also to low molecular weight fragments, again suggesting that not all copies of *PRT1* are subtelomeric.

The *P.carinii PRT1* gene family shows some striking similarities to that of *MSG*. Both are composed of many genes, copies of which are found on most *P.carinii* chromosomes and show sequence heterogeneity. Some copies of *PRT1* are contiguous with *MSG* and are located in the subtelomeric regions of the *P.carinii* chromosomes.

It is interesting to note that one of the major components of the cell surface of *Leishmania* has proteolytic activity. The *Leishmania* major surface protease (*msp* or *gp63*), a zinc endoprotease, is found in all species of *Leishmania* and is encoded by a family of genes, some of which are tandemly arrayed (Bouvier *et al.*, 1989; Webb *et al.*, 1991). Expression of different copies of the gene is regulated during the development of the parasite and different isoforms of the protein are found in the promastigote stage in the gut of the sand fly and in the amastigote stage in the phagolysosomes of the macrophages (Frommel *et al.*, 1990; Roberts *et al.*, 1995; Ramamoorthy *et al.*, 1995). The major surface protease is thought to play an important role in the virulence of *Leishmania* by involvement in the degredation of components of the extracellular matrix and by facilitating promastigote attachment to host macrophages (McMaster *et al.*, 1994). Immunisation with MSP protein confers partial protection of mice against *Leishmania* infection (Abdelhak *et al.*, 1995).

The proteins encoded by the *P.carinii PRT1* gene family show highest homology to the subtilisin-like serine proteases. A wide diversity of different types of precursor proteins are processed by this family of proteases to mature and active regulatory proteins, but the precise function of many of these proteases has not yet been determined. Some of the fungal homologues have been shown to function in the processing of several proteins, such as the S. *cerevisiae KEX*2 gene product which processes both the pheromone α-factor and the killer toxin (Fuller *et al.*, 1989). The *krp* gene product from *S.pombe*, which cleaves the pheromone precursor pro-P-factor to its active form, is thought to also function in the processing of other regulatory proteins, since its activity is essential for cell viability (Davey *et al.*, 1994). The *XPR6* gene product from *Y. lipolytica*, although not essential for cell viability, when disrupted was found to cause aberrant growth and morphology (Enderlin and Ogrydziak, 1994). The function of the products of the *P.carinii PRT1* gene family is not yet understood but it is likely to play an important role in the life cycle and possibly also the pathogenicity of the organism.

Identification and sequencing of a *PRT1* gene from *P.carinii* sp. f *hominis*

PCR strategies using degenerate primers designed using *P.carinii* sp. f. *carinii PRT1* sequence information failed to isolate any *P.carinii* sp. f. *hominis PRT1* clones. The strategies employed included single round PCR and nested PCR, on post mortem samples from infected patients.

Given the failure of these approaches, it was decided to try to obtain additional sequence data from *P.carinii* derived from other organisms.

MATERIALS AND METHODS

Samples

Samples of *Pneumocystis carinii* sp. f. *hominis* were derived from HIV positive patients by fibreoptic bronchoscopy, an aliquot of this bronchoscopic alveolar lavage (BAL) sample being immediately frozen, stored at -20°C and transported to the Institute of Molecular Medicine for DNA extraction (samples D503B and D122B). One sample (C180) was derived from a post mortem lung from an HIV-negative patient; the parasites were first enriched by successive filtration through 70 μm, 12 μm and 8μm filters.

Samples of *Pneumocystis* from the infected lungs of four other mammalian hosts were used. These were *Pneumocystis carinii* sp. f. *muris* (mouse derived), *Pneumocystis carinii* sp. f. *mustelae* (ferret derived), *Pneumocystis carinii* sp. f. *suis* (pig derived), *Pneumocystis carinii* sp. f. *carinii* (rat-derived) and *Pneumocystis carinii* sp. f. *rattus* (rat derived). These were enriched for parasites prior to DNA extraction.

DNA Extraction

DNA was extracted from an enriched parasite preparation by proteinase K digestion, followed by phenol-chloroform extraction. The DNA was purified and concentrated using a DNA binding resin (Promega Wizard DNA Clean-UP System).

DNA Amplification

In general the following conditions were used in all PCR reactions. The final concentration of the reaction mix was 50mM KCl, 10mM Tris (pH 8.0), 0.1% Triton X-100, 3mM $MgCl_2$, 400μM of each deoxynucleoside triphosphate, 1μM of each oligonucleotide primer and 0.025U of *Taq* polymerase (Promega) per ml. A total of forty cycles was used with 10 cycles at 94°C for 1.5 min (denaturation), annealing at a temperature between 48°C and 55°C dependant on primer Tm and required stringency of reaction for 1.5min and 72°C for 2min (extension), followed by 30 cycles at 94°C for 1.5min, 63°C for 1.5min and 72°C for 2min (the increased temperature at annealing now including the *Eco*R1 site at the 5' end of the primers). Where there was no *Eco*R1 site in the primer or where particularly low stringency was required all 40 cycles were carried out at the lower annealing temperature. A positive control of rat *Pneumocyctis* DNA (rat 1458 or rat 1189) was included in each PCR reaction. Negative controls of no added template DNA were included after each sample to monitor for cross contamination. In later PCR reactions, when degenerate primers were being used, a negative control of human DNA (Sigma), at a final concentration of 0.8ng/µl, was included to monitor for non-specific amplification of human DNA, which was unavoidably co-extracted with all human *Pneumocystis* DNA samples. The primers used are shown in Table 1 herein (and Table 1 of Lugli *et al* 1997)..

All PCR products were electrophoretically separated out on 1.2% or 1.5% agarose gels containing ethidium bromide, visualised under ultraviolet light.

Determination of the complete sequence of a copy of *P.carinii* sp. f. *hominis PRT1* gene

A number of different approaches are available for the isolation of the complete gene sequence of a *P.carinii* sp. f. *hominis PRT1* gene. Some of the possible approaches are described below in detail.

DNA and RNA is prepared from *P.carinii* sp. f. *hominis* organisms, obtained from either bronchoalveolar lavage samples from *P.carinii* infected patients or from post-mortem lung samples.

i) *P.carinii* sp. f. *hominis* genomic library

A *P.carinii* sp. f. *carinii* genomic library is constructed in λFIX and this is screened with the cloned fragment of *PRT1*. Positive recombinant phage are analysed by further rounds of screening, and full length clones selected for analysis. The arrangement of introns within the gene sequence is determined. The genomic organisation of copies of *PRT1* is elucidated, and in particular the relationship with gene copies of MSG. The chromosomal organisation of different *PRT1* copies is examined, including the analysis of copies which are in the subtelomeric regions and others which are at an internal location.

ii) Expressed copies of *PRT1*

Two different approaches can be used to examine transcribed copies of *PRT1*. In the first, Random Amplification of cDNA Ends (RACE) is used to extend 5'- and 3'- of the cloned fragment of *PRT1*, using total RNA or poly A^+ RNA from the enriched parasite preparation. Primers are designed to the sequence of the cloned fragment for use in this technique. The second approach is the construction of a cDNA library in λZAP from *P.carinii* sp. f. *hominis*, which is then screened with the cloned fragment. Different recombinant clones are compared for variation in sequence and used for expression studies.

Expression

i) Expression of cloned fragment of *P.carinii* sp. f. *hominis* *PRT1* (H13)

The known portion of the catalytic domain is subcloned into the pET32a expression vector and expressed in an *E. coli* expression system. Recombinant protein is purified and used to raise polyclonal antiserum in rabbits. In addition, synthetic peptides designed to the PRT1 derived amino acid sequence are used in the production of antibodies.

ii) Expression of the complete gene sequence and fragments of the gene spanning different domains.

Recombinant protein is expressed and purified from different domains and from the complete sequence, for use in the production of antibodies, and in biochemical and immunohistochemical studies.

Biochemical studies

Biochemical studies are performed to determine the substrate specificity of the protease and the optimum conditions (e.g. pH, metal cofactors) for proteolytic activity. This provides an *in vitro* system for the testing of inhibitors to the *PRT1* protease. Crystallisation of the recombinant protein is carried out and the 3-D structure of the protein determined by X-ray crystallography and compared with the 3D structure of the four other subtilisin-like serine proteases whose structure has previously been determined. These structural data can used for purposes including the design of specific inhibitors of *PRT1*, and the prediction of antigenically important epitopes.

Immunohistochemistry

Antibodies raised to the recombinant *PRT1* protein or to synthetic peptides can be used in the analysis of the subcellular localisation of *PRT1* in *P.carinii* organisms, using both light microscopy and electron microscopy with immunogold.

Table 1

Oligonucleotide primers

Primer	Sequence
Pcprot1d/R1	GGGAATTCTA$^{T}_{C}$$^{T}_{A}$$^{C}_{G}NTG^{T}_{C}$$^{T}_{A}$$^{C}_{G}$NTGGGGNCC
Pcprot16d/RI	GGGAATTCCA$^{G}_{T}$GgiACi$^{G}_{A}$GiTG$^{T}_{C}$GCiGG
Pcprot17d/RI	GGGAATTCA$^{C}_{T}$$^{G}_{A}Tci^{T}_{C}$$^{G}_{T}$CCAiGTiA$^{G}_{A}$$^{G}_{A}T^{T}_{C}$iGG
Pcprot18d/RI	GGGAATTCTAiGC$^{G}_{A}$TciAi$^{T}_{C}$TTiCC$^{A}_{G}$$^{A}_{TA}$iCC
Pcprot24d/RI	GGGAATTC$^{G}_{A}$CC$^{A}_{C}$GAATA$^{T}_{C}$GTAGAAGC
Pcprot25d/RI	GGGAATTCGTTTT$^{T}_{C}$GG$^{G}_{A}$$^{A}_{T}$$^{C}_{G}$$^{A}_{TC}$GAGG$^{A}_{T}$GG
Pcprot26d/RI	GGGAATTC$^{A}_{T}$GCAA$^{T}_{G}$AGGT$^{A}_{G}$$^{T}_{C}$$^{A}_{G}$GAAGCAGA
Pcprot31/RI	GGGAATTCGAAGATGTTGATATTGAGGAG
Pcprot32/RI	GGGAATTCATCGTCTCTTATCGCACCC
Pcprot33/RI	GGGAATTCTCAACTCAACTAATACC
Pcprot39/RI	GGGAATTCAGGAATGATTTTTGTGGGCT
73jEx4/RI	GGGAATTCTTATGGAACAGCTGTTTCC
73jEx5/RI	GGGAATTCATCAATAGACTCTCCG
PcprotH34/RI	GGGAATTCTTGCGAATATTATCCGGGC
PcprogH35/RI	GGGAATTCGCACTTCCACCTGCATATG

Oligonucleotide Sequences. Note that I = inosine and N = any base in degenerate sequences. The oligonucleotides above have SEQ ID NOS: 1-15, according to the order in which they appear in the above table.

Single round PCR on Rat Variant, Mouse, Ferret and Pig derived *P.carinii*

Single round PCR on *P.carinii* sp. f. *rattus* and *P.carinii* sp.f. *muris* samples gave strong amplification products at the same Mr as the rat *P.carinii* positive control. Primers used were Pcprot1/R1 and Pcprot3/R1. Sequence data is shown in Figure 2.

Single Round PCR on Human Post Mortem Sample using Redesigned Primer

New primers were designed based on regions of homology of the newly obtained rat variant *P. carinii* and mouse *P. carinii PRT1* sequences with the rat prototype *P. carinii* sequence at both the DNA level and amino acid level. These were not fully degenerate, given that *Pneumocystis* DNA shows a high AT bias (60-70%); unless the sequence data suggested otherwise only A or T was used at potentially degenerate sites (as seen in the amino acid sequences). These new primers were used in reactions with one another and previously used primers. Of these reactions, only Pcprot16d/R1 and Pcprot26d/R1 gave a clear positive product at the expected Mr, close to that of the rat *P. carinii* positive control (~600 b.p.). The primers used were Pcprot25d/R1 + Pcprot26d/R1; Pcprot1d/R1 + Pcprot26d/R1; Pcprot16d/R1 + Pcprot26d/R1; Pcprot25d/R1 + Pcprot17d/R1; Pcprot25d/R1 + Pcprot18d/R1; Pcprot25d/R1 + Pcprot24d/R1. The PCR products from the reactions were cloned and sequenced. Of the clones sequenced one contained an insert which showed homology to the *PRT1* gene. Sequence data over the catalytic domain is shown in Figures 2 and 3.

	Mt LSU rRNA	mt SSU rRNA	arom (DNA)	arom (aa)	PRT1 (DNA)	PRT1 (aa)
Variant Rat *P. carinii*	13	12	-	-	28-31	49-53
Mouse *P. carinii*	14	8	7	7	27-28	43-46
Human *P. carinii*	24	18	18	20	42	67

Table showing percentage divergence of prototype rat-derived Pneumocystis (*P.carinii* sp. f. *carinii*). mt LSU rRNA - mitochondrial large subunit rRNA; mt SSU rRNA - mitochondrial small subunit rRNA. Values for Variant rat *P. carinii* from two clones; values for Mouse *P. carinii* from three clones. DNA divergence calculated with Jukes-Cantor correction method. Protein divergence calculated using Kimura protein distance.

The above table shows that the *PRT1* gene differs between *P.carinii* from different host organisms by far more than many other genes so far studied. Thus in *P.carinii* sp. f. *hominis* the *PRT1* DNA sequence is around twice as divergent from *P.carinii* sp. f. *carinii* compared to other sequences and the amino acid sequence is over three times as divergent as the *arom* sequence. This is even more striking given that the *PRT1* data are taken from the catalytic domain which should contain the highest level of conservation (catalytic, substrate binding, oxyanion hole and disulphide bridge residues). A similar level of divergence has previously been observed in the *MSG* (also called Glycoprotein A; *gpA*) genes. Indeed, early attempts to amplify some portions of *gpA/MSG* from *P.carinii* sp. f. *hominis* by PCR using primers based on the *P.carinii* sp. f. *carinii* sequence failed (Kovacs *et al.*, 1993; Wright *et al.*, 1994).

A high level of divergence is also seen in the *PRT1* sequences from *P.carinii* sp. f. *rattus* and *P.carinii* sp. f. *muris* where the

PRT1 DNA sequences are two to four times as divergent as the other sequences and the mouse *P. carinii PRT1* amino acid sequence is over six times more divergent than that of *arom*.

The homology of the amino acid sequences from all three types of *Pneumocystis* to the subtilisin-like serine proteases is high. Of the known conserved residues, most can be seen to be conserved in the *PRT1* sequences (where the data are available). Certainly in the *P.carinii* sp. f. *hominis PRT1* amino acid sequence there is greater conservation of the negatively charged amino acids at the substrate-binding face than is seen in the *P.carinii* sp. f. *carinii* sequence. Although the homology to the subtilases is unmistakable, there is considerable variation to be seen between the *PRT1* sequences. This presumably reflects differences in substrate specificity, whether the substrate is a host protein (or proteins) or a parasite protein (e.g. gpA/MSG).

The function of the subtilisin-like serine proteases so far studied is in the specific endoproteolytic processing of precursor proteins to their active form. Although the precise function of many subtilases is yet to be determined, some fungal homologues have been shown to be vital to cell viability or normal function. Thus *krp* in *S. pombe* has been shown to be vital to cell viability and disruption of *XPR6* in *Y. lipolytica* causes aberrant growth and morphology. Parallels may also be drawn between *Gp63* in *Leishmania* and *PRT1* in *Pneumocystis*, as discussed in the introduction. The functions of the PRT1 proteins are not yet fully established, but it seems likely to be important to the life-cycle and/or the pathogenesis of the organism. The cloning of this gene, most especially from *P.carinii* sp.f. *hominis*, is thus a step towards the design of an effective anti-*Pneumocystis* drug.

Generation of anti-PRT1 antibodies

Polyclonal antiserum was generated in rabbits to synthetic peptides, designed to the *Pneumocystis carinii* sp. f. *carinii PRT1* sequence. Regions of the protein which were likely to be immunogenic were predicted using the appropriate software, and peptides (15 mers) to six different regions were synthesized. A mixture of six synthetic peptides was administered by subcutaneous injection to rabbits (New Zealand white). An antibody response was elicited by standard procedures, using Freunds complete adjuvant for the first injection and Freunds incomplete adjuvant for subsequent injections.

The resulting polyclonal antisera were tested against the peptides. The greatest cross-reactivity of the antisera was found with Peptide 7, designed to a region of the catalytic domain (amino acid residues 424 - 438 of the PRT1(73j) sequence) and with Peptide 9, designed to the pro-domain (amino acid residues 64 - 78 of the PRT1(73j) sequence).

Peptide sequences

TWRDVQALIVETAVP	(2)	(SEQ ID NO: 16)
ITSPSGVTSVLAHRR	(4)	(SEQ ID NO: 17)
ESEGVPPPSYPFLSR	(5)	(SEQ ID NO: 18)
ASTPLAAGVIALLLS	(7)	(SEQ ID NO: 19)
FRGESIVGNWTIDVE	(8)	(SEQ ID NO: 20)
DNQHIFSIEKGVLED	(9)	(SEQ ID NO: 21)

EXAMPLES

Example 1

Expression of portions of the rat-derived *P. carinii* (*P. carinii* sp. f. *carinii*) PRT1(73j) gene.

The *E. coli* expression vector pET32a (Novagen, Madison, WI) was used. This vector contains an inducible T7lac promotor, a 6-His tag, a multiple cloning site and the recombinant protein is expressed as fusion protein with the Trx-tag thioredoxin protein (109 amino acids).

Recombinant thioredoxin fusion proteins are generally more soluble and remain in the *E. coli* cytoplasmic fraction. Three different regions of the *PRT1(73j)* gene were cloned into pET32a: i) Cat2f1, a portion of the catalytic domain, 585bp in length, from base 790 to base 1375; ii) F1a1j, a portion of the pro-domain, 255bp in length, from base 120 to base 375; iii) G1b1c, a portion of the P domain, 384 bp in length, from base 1515 to base 1899.

The specific fragments were amplified by PCR from the PRT1(73j) sequence as follows - i) Cat2f1 using primers Pcprot39/R1 and 73j Ex4; ii) F1a1j using primers Pcprot31/RI and Pcprot32/RI; iii) G1b1c using primers Pcprot33/RI and 73jEx5/RI (see Table 1). All primers included an EcoRI site the 5′ end to facilitate cloning. The fragments were initially cloned into the plasmid vector pUC, linearized with *Eco*RI and treated with alkaline phosphatase, to produce a stable, high copy number, recombinant plasmid. The recombinant DNA was then subcloned into the *Eco*RI site of the expression vector pET32a.

2. Transformation of *E. coli* with recombinant plasmids

E. coli DH5α competent cells were transformed with the recombinant plasmids. The cells were transformed with recombinant pUC plasmids, and also recombinant pET32a plasmids. The recombinant expression vector pET32a constructs were also transferred into *E. coli* DE3 (BL21) cells, for expression of the recombinant peptides.

3. Expression of recombinant PRT1 polypeptides

The recombinant pET32a constructs, transformed into *E. coli* DE3(BL21) were induced with IPTG, and the bacteria were grown for 3 to 4 hours. The cells were collected by centrifugation and disrupted by sonication. The bacterial proteins were separated by SDS-PAGE and electrophoretically transferred to nitrocellulose filter. The immobilised proteins were cross-reacted with anti-thioredoxin antibody (Sigma), and the bound antibody was visualised with a swine anti-rabbit immunoglobulins secondary antibody, conjugated to alkaline phosphatase. A band of the expected size (24kDa) was seen in the control vector pET32a, (lane 1) corresponding to the thioredoxin fusion protein and the His-tag. Bands corresponding to the expected sizes of the recombinant PRT1 protein fragments were observed (Figure 7, lanes 2 and 3).

4. Preparation of polyclonal mono-specific antibodies

Polyclonal antisera raised against the six synthetic peptides were affinity purified. The peptide (Peptide 7 or Peptide 9) was covalently linked to an amine reactive support. Immunoglobulins which cross-reacted to the peptide were specifically retained by the column, and subsequently eluted. In this way, two polyclonal mono-specific antibodies were produced, anti-Peptide 7 and anti-Peptide 9.

5. Cross-reactivity of polyclonal, mono-specific antibodies with recombinant PRT1 polypeptides

Expressed proteins from transformation of *E. coli* DE3(BL21) with recombinant expression vector to the pro-domain (F1a1j) or to the catalytic domain (Cat2f1) were separated by SDS-PAGE and electrophoretically transferred to nitrocellulose membrane. The anti-Peptide 7 mono-specific antibody was shown to cross-react with the recombinant Cat2f1 polypeptide, but not to F1a1j or to the protein produced by the control plasmid pET32a. Likewise, the anti-Peptide 9 antibody specifically cross-reacted with the F1a1j polypeptide. These results confirm the specificity of the mono-specific antisera to the two distinct domains of the PRT1 protein.

6. Identification of PRT1 protein in *P.carinii* sp. f. *carinii* organisms

P.carinii sp. f. *carinii* organisms were extracted and enriched from infected rat lungs. Organisms were disrupted by heating to 95°C in denaturing solution and the proteins separated by SDS-PAGE, followed by transfer to nitocellulose filters. The immolbilised proteins were cross-reacted with the anti-Peptide 7 and the anti-Peptide 9 antibody. Bound antibody was detected using an anti-rabbit secondary antibody, conjugated to alkaline phosphatase. A single, major band, at 40 kDa, was seen with each of the mono-specific antibodies. In addition, another major band at 38 kDa was seen with anti-Peptide 7 antibody and minor bands at 98 kDa and 16 kDa. With the anti-Peptide 9 antibody, minor bands at 200kDa, 98kDa and 43 kDa were observed. The predicted size of the full length PRT1 protein ranges from 87 to 102 kDa. The proteins detected with the mono-specific antibodies are assumed to be the products of autocatalysis at a number of dibasic residues found in the PRT1 sequence.

7. Sub-cellular localisation of the PRT1 protein in *P.carinii* sp. f. *carinii* organisms

Sections of *P.carinii* sp. f. *carinii* infected rat lungs, formalin fixed and embedded in paraffin, were prepared and incubated with anti-Peptide 7 antibody. Bound antibody was detected using a swine anti-rabbit immunoglobulin secondary antibody, conjugated to horse radish peroxidase, and the organisms viewed by light microscopy. The specific distribution of the antibody on the *P.carinii* sp. f. *carinii* organisms was characteristic of surface localisation of the PRT1 protein in the organisms.

Example 2

Expression of a portion of the human-derived *P. carinii* (*P. carinii* sp. f. *hominis*) *PRT1* gene

1. Construction of recombinant vector containing a portion of the *P.carinii* sp. f. *hominis PRT1* gene

The *E.coli* expression vector pET32a (Novagen, Madison, WI) was used. This vector contains an inducible T7lac promotor, a 6-His tag, a multiple cloning site and recombinant protein is expressed as fusion protein with the Trx-tag thioredoxin protein (109 amino acids). Thioredoxin fusion proteins are generally more soluble and remain in the *E.coli* cytoplasmic fraction.

A 367bp portion of the cloned *P. carinii* sp. f. *hominis PRT1(H13)* sequence was amplified using PCR with the primers PcprotH34/RI and PcprotH35/RI, corresponding to position 111 to position 478 on the PRT1 (H13) sequence, in the catalytic domain of the gene (see Table 1). The primers included an *Eco*RI site at the 5′ end to facilitate cloning. The resulting fragment (H1a1a) was initially cloned into the *Eco*RI site of the plasmid vector pUC, and then subcloned into the *Eco*RI site of the expression vector pET32a.

2. Transformation of *E. coli* with recombinant plasmids

E. coli DH5α competent cells were transformed with the recombinant plasmid. The cells were transformed with the recombinant pUC plasmid, and also the recombinant pET32a plasmid. The recombinant expression vector pET32a construct was also transferred into *E. coli* DE3 (BL21) cells, for expression of the recombinant peptide.

3. Expression of recombinant *P.carinii* sp. f. *hominis* PRT1 peptide

The recombinant pET32a construct (H1a1a), transformed into *E. coli* DE3(BL21) was induced with IPTG, and the bacteria were grown for 3 to 4 hours. The cells were collected by centrifugation and disrupted by sonication. The bacterial proteins were separated by SDS-PAGE and electrophoretically transferred to nitrocellulose filter. The immobilised proteins were cross-reacted with anti-thioredoxin antibody (Sigma), and the bound antibody was visualised with a swine anti-rabbit immunoglobulins secondary antibody, conjugated to alkaline phosphatase. A band of the expected size (24kDa) was seen in the vector pET32a control, (lane 1) corresponding to the thioredoxin fusion protein and the His-tag. A band corresponding to the expected size of the recombinant *P.carinii* sp. f. *hominis* PRT1 protein fragment was observed (Figure 7, lane 4).

4. Identification of PRT1 protein in *P.carinii* sp. f. *hominis* organisms

P.carinii sp. f. *hominis* organisms were extracted from bronchoalveolar lavage fluid from a patient with *P. carinii* pneumonia. The organisms were disrupted by heating to 95°C in denaturing solution and the proteins separated by SDS-PAGE, followed by transfer to nitrocellulose filters. The immobilised proteins were cross-reacted with the anti-Peptide 7 and the anti-Peptide 9 antibody. Bound antibody was detected using an anti-rabbit secondary antibody, conjugated to alkaline phosphatase. Two major bands, at 56 kDa and 49 kDa was seen with each of the mono-specific antibodies. In addition, minor bands at 116kDa, 95kDa, 86 kDa and 39 kDa were seen with the anti-Peptide 7 antibody, and at 200 kDa, 116kDa, 95kDa, 86 kDa and 29 kDa with the anti-Peptide 9 antibody. The proteins detected with the mono-specific antibodies are assumed to be the products of autocatalysis at a number of dibasic residues found in the *P.carinii* sp. f. *hominis* PRT1 sequence.

REFERENCES

Abdelhak, S., Louzir, H., Timm, J., Blel, L., Banlasfar, Z., Lagranderie, M., Gheorghiu, M., Dellagi, K. & Gicquel, B. (1995). Recombinant BCG expressing the leishmania surface antigen Gp63 induces protective immunity against Leishmania major infection in BALB/c mice. Microbiology **141**, 1585-1592.

Banerji, S., Wakefield, A.E., Allen, A.G., Maskell, D.J., Peters, S.E. and Hopkin, J.M. (1993). The cloning and characterization of the arom gene of Pneumocystis carinii. J Gen Microbiol **139**, 2901-2914.

Betzel, C., Pal G.P. and Saenger W. (1988). Three-dimensional structure of proteinase K at 0.15nm resolution. Eur J Biochem **178**, 155-171.

Bott, R., Ultsch, M., Kossiakoff, A., Graycar, T., Katz, B. and Power, S. (1988). The three-dimensional structure of Bacillus amyloliquefaciens subtilisin at 1.8 Å and an analysis of the structural consequences of peroxide inactivation. J Biol Chem **263**, 7895-7906.

Bouvier, J., Bordier, C., Vogel, H., Reichelt, R. & Etges, R. (1989). Characterization of the promastigote surface protease of Leishmania as a membrane-bound zinc endopeptidase. Mol & Biochem Paras **37**, 235-246.

Cushion, M.T., Kaselis, M., Stringer, S.L. and Stringer, J.R. (1993). Genetic stability and diversity of Pneumocystis carinii infecting rat colonies. Infect Immun **61**, 4801-4813.

Davey, J., Davis, K., Imai, Y., Yamamoto, M., and Matthews, G. (1994). Isolation and characterization of krp, a dibasic endopeptidase required for cell viability in the fission yeast Schizosaccharomyces pombe. EMBO Journal **13**, 5910-5921.

Dyer, M., Volpe, F., Delves, C.J., Somia, N., Burns, S. and Scaife, J.G. (1992).Cloning and sequence of a β-tubulin cDNA from Pneumocystis carinii: possible implications for drug therapy. Mol Microbiol **6**, 991-1001.

Enderlin, C. S., and Ogrydziak, M. (1994). Cloning, nucleotide sequence and functions of XPR6, which codes for a dibasic processing endoprotease from the yeast Yarrowia lipolytica. Yeast **10**, 67-79.

Frommel, T. O., Button, L. L., Fujikura, Y. & McMaster, W. R. (1990). The major surface glycoprotein (GP63) is present in both life stages of Leishmania. Mol & Biochem Paras **38**, 25-32.

Fuller, R. S., Brake, A., and Thorner, J. (1989). Yeast prohormone processing enzyme (KEX2 gene product) is a Ca^{2+}-dependent serine protease. Proc. Natl. Acad. Sci. USA **86**, 1434-1438.

Garbe, T. R. and Stringer, J. R. (1994). Molecular characterization of clustered variants of genes endocing major surface antigens of human Pneumocystis carinii. Infect Immun **62**, 3092-3101.

Gigliotti F. (1992). Host species-specific antigenic variation of a mannosylated surface glycoprotein of Pneumocystis carinii. J Infect Dis **165**, 329-336.

Gros, P., Betzel, C., Dauter, Z., Wilson, K. S and Hol, W. G. J. (1989). Molecular dynamics refinement of a thermitase-eglin-c-complex at 1.98 Å resolution and comparison of two crystal forms that differ in calcium content. J Mol Biol **210**, 347-367.

Haidaris, P. J., Wright, T. W., Gigliotti, F. and Haidaris, C. G. (1992). Expression and characterization of a cDNA clone encoding an immunodominant surfact glycoprotein of Pneumocystis carinii. J Infect. Dis **166**, 1113-1123.

Hirono, S., Akagawa, H., Iitaka, Y. and Mitsui, Y. (1984). Crystal structure at 2.6 Å resolution of the complex of subtilisin BPN with Streptomyces subtilisin inhibitor. J Mol Biol **178**, 389-414.

Klemm, U., Müller-Esterl, W. and Engel, W. (1991). Acrosin, the peculiar sperm-specific serine protease. Human Genetics **87**, 635-641.

Kovacs, J. A., Powell, F., Edman, J. C., Lundgren, B., Martinez, A., Drew, B. and Angus, C. W. (1993). Multiple genes encode the major surface glycoprotein of Pneumocystis carinii. J Biol Chem **268**, 6034-6040.

Lugli, E.B. and Wakefield, A.E. (1996). A novel subtelomeric multi-gene family in *Pneumocystis carinii* . 4th International Workshop on Opportunistic Protists, Tuscon, Arizona, USA, June 1996.

Lugli, E.B.,Allen, A.G. and Wakefield, A.E. (1997) A *Pneumocystis carinii* multi-gene family with homology to subtilisin-like serine proteases. Microbiology **143**: 2223-2236.

Lundgren, B., Lipschik, G.Y. and Kovacs, J.A. (1991). Purification and characterization of a major human Pneumocystis carinii surface antigen. J Clin Invest **87**, 163-170.

McMaster, R. R., Morrison, C. J., MacDonald, M.H. & Joshi, P. B. (1994). Mutational and functional analysis of the Leishmania surface metalloproteinase GP63 : similarities to matrix metalloproteinases. Parasitology **108**, S29-S36.

McPhalen, C.A. and James, M. N. G. (1988). Structural comparison of two serine proteinase-protein inhibitor complexes: eglin-c-subtilisin Carlsberg and CI-2-subtilisin Novo. Biochem **27**, 6582-6598.

Peters, S.E., Wakefield, A.E., Banerji, S. and Hopkin, J.M. (1992). Quantification of the detection of Pneumocystis carinii by DNA amplification. Molec Cell Probes **6**, 115-117.

Ramamoorthy, R., Swihart, K. G., McCoy, J. J., Wilson, M. E. & Donelson, J. E. (1995). Intergenic regions between Tandem gp63 genes influence the differential expression of gp63 RNAs in Leishmania chagasi promastigotes. J Biol Chem **270(20)**, 12133-12139.

Roberts, S. C., Wilson, M. E. & Donelson, J. E. (1995). Developmentally regulated expression of a novel 59-kDa product of the major surface protease (Msp or gp63) gene family of Leishmania Chagasi. J Biol Chem **270(15)**, 8884-8892.

Roditi, I., Schwarz, H., Pearson, T.W., Beecroft, R.P., Liu, M.K., Richardson, J.T., Buhring, H.J., Pleiss, J., Bulow, R., Williams, R.O. and Overath, P. (1989). Procyclin gene expression and loss of the variant surface glycoprotein during differentiation of Trypanosoma brucei. J cell Biol **108**, 737-746.

Roebroek, A. J. M., Creemers, J. W. M., Pauli, I. G. L., Kurzik-Dumke, U., Rentrop, M., Gateff, E. A. F., Leunissen, J. A. M. and Van de Ven, W. J. M. (1992). Cloning and functional expression of Dfurin2, a subtilisin-like proprotein processing enzyme of Drosophila melanogaster with multiple repeats of a cysteine motif. J Biol Chem **267**, 17208-17215.

Rogers, J. (1985). Exon shuffling and intron insertion in serine protease genes. Nature **315**, 458-459.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). Molecular cloning: a laboratory manual. Cold Spring Habour, New York: Cold Spring Habour Laboratory Press.

Siezen, R. J., de Vos, W. M., Leunissen, J. A. M. and Dijkstra, B. W. (1991). Homology modelling and protein engineering strategy of subtilases, the family of subtilisin-like serine proteinases. Protein Eng **4**, 719-737.

Stringer, S. L., Garbe, T., Sunkin, S. M. and Stringer, J. R. (1993). Genes encoding antigenic surfact glycoproteins in Pneumocystis from humans. J Euk Microbiol **40**, 821-826.

Sunkin, S. M., and Stringer, J. R. (1996). Translocation of surface antigen genes to a unique telomeric expression site in Pneumocystis carinii. Mol Microbiol **19**, 283-295.

Sunkin, S. M., Stringer, S. L. and Stringer, J. R. (1994). A tandem repeat of rat-derived Pneumocystis carinii genes encoding the major surface glycoprotein. J Euk Microbiol **41**, 292-300.

Tanguy-Rougeau, C., Wesolowski-Louvel, M. and Fukuhara, H. (1988). The Kluyveromyces lactis KEX1 gene encodes a subtilisin-type serine proteinase. FEBS Lett. **234**, 464-470.

Teplyakov, A. V., Kuranova, I. P., Harutyunyan, E. H. and Vainshtein, B. K. (1990). Crystal structure of thermitase at 1.4 Å resolution. J Mol Biol **214**, 261-279.

Underwood, A. P., Louis, E.J., Borts, R. H., Stringer, J. R. and Wakefield, A. E. (1996). Pneumocystis carinii telomere repeats are composed of TTAGGG and the subtelomeric sequence contains a gene encoding the major surface glycoprotein. Mol. Microbiol **19**, 273-281.

van den Ouweland, A.M.W., van Duijnhove, H.L.P., Keizer, G.D., Dorssers, L.C.J. and Van de Ven, W.J.M. (1990). Structural homology between the human fur gene product and the subtilisin-like protease encoded by yeast KEX2. Nucl Acids Res **18**, 664.

Van de Ven, W. J. M. and Roebroek, A. J. M. (1993). Structure and function of eukaryotic proprotein processing enzymes of the subtilisin family of serine proteases. Critical Rev in Oncogenesis **4**, 115-136.

Volpe, F., Dyer, M., Scaife, J. G., Derby, G., Stammers, D.K. and Delves, C. J. (1992). The multifolate folic acid synthesis fas gene of Pneumocystis carinii appears to encode dihydropteroate synthase and hydroxymethyldihydropterin pyrophosphokinase. Gene **112**, 213-218.

Volpe, F., Ballantine, S. P., and Delves, C. J. (1993). The multifunctional folic acid synthesis fas gene of Pneumocystis carinii encodes dihydroneopterin aldolase, hydroxymethyldihydropterin pyrophosphokinase and dihydropteroate synthase. Eur J Biochem **216**, 449-458.

Wada, M., and Nakamura, Y. (1994). MSG gene cluster encoding major cell surface glycoproteins of rat Pneumocystis carinii. DNA Research **1**, 163-168.

Wada, M., Kitada, K., Saito, M., Egawa, K. and Nakamura, Y. (1993). cDNA sequence diversity and genomic clusters of major surface glycoprotein genes of Pneumocystis carinii. J Infect Dis **168**, 979-985.

Wada, M., Sunkin, S.M., Stringer, J.R. and Nakamura, Y. (1995). Antigenic variation by positional control of major surface glycoprotein gene expression in Pneumocystis carinii. J Infect Dis **171**, 1563-1568.

Webb, J. R., Button, L. L. & McMaster, W. R. (1991). Heterogeneity of the genes encoding the major surface glycoprotein of Leishmania donovani. Mol & Biochem Paras **48**, 173-184.

Wright, T. W., Simpson-Haidaris, P. J., Gigliotti, F., Harmsen, A. G. & Haidaris, C. G. (1994). Conserved sequence homology of cysteine-rich regions in genes encoding glycoprotein A in Pneumocystis carinii derived from different host species. Inf & Immun **62**, 1513-1519.

Wright, T. W., Bissoondial, T. Y., Hairdaris, C. G., Gigliotti, F. & Simpson Haidaris, P. J. (1995). Isoform diversity and tandem duplication of the glycoprotein A gene in ferret Pneumocystis carinii. DNA Research **2**, 77-88.

Zhang, J. and Stringer, J.R. (1993). Cloning and characterization of an alpha-tubulin-encoding gene from rat-derived Pneumocystis carinii. Gene 123,137-141.

Figure Legends

Figure 2

Nucleotide sequence alignments of part of the catalytic domain of *PRT1.* 1-3 page, 11-3-73j andd 1-3prp5e from *P. carinii* f.sp. *carinii* [(8)]; *ratv5prt1 and ratv16prt1 from P. carinii* f. sp. rattus; mousee1prt1, mouse7prt1 and mouse13prt1 from *P. carinii* f. sp. *muris*; humanprt1 from *P. carinii* f. sp.

Figure 3

Amino acid sequence alignments of part of the catalytic domain of *PRT1*, translated from the nucleotide sequences (Figure 2). Pagaprt1, 73jpart1 and prp5eprt1 from *P. carinii* f. sp. *P. carinii*[(8)]; ratv5prt1 and ratv16prt1 from *P. carinii* f. sp. *rattus*; mouse1prt1, mouse7prt1 and mouse13part1 from *P. carinii* f. sp. *muris*; humanprt1 from *P. carinii* f. sp. *hominis*. ⇓ marks conserved amino acids; numbering according to full amino acid sequence of cDNA clone 73j[(8)]; an asterisk marks positions of charge conservation in subtilases (see text).

Figure 4

Alignment of the *P.carinii* sp. f. *carinii* PRT1 deduced amino acid sequences from the genomic clone Paga, the cDNA clone 73j and the three overlapping PCR products amplified from a cDNA library corresponding to the 5' region (Prp5e), the central region (M14), and the 3' region (Prp2g). The deduced amino acid sequences of PCR products amplified from five different regions of the *PRT1* gene family were also aligned; the catalytic domain: Prp1a, Prp3a, Prp7a; the boundary of the catalytic domain and the P-domain: Prp2c, Prp3c, Prp4c; the P-domain: Prptaf2, Prpf4, Prp5f; the proline-rich region: Pcr-19, Pcr-14, Pcr-5, Pcr-3, Pcr-1, Lam-1; the C-terminal region: Prpg4, Prpg3, Prp5g. Gaps were introduced to maximize homology; identical amino acids are boxed.

Figure 6

Schematic representation of the *P. carinii* sp. f. *carinii* PRT1. Patterned boxes represent different domains; small dots represent hydrophobic regions (HR), diagonal lines indicate the catalytic domain (CAT), woven pattern indicates the P-domain (P), vertical lines indicate the proline-rich region, squares indicate the serine-threonine rich region (STR). Boxes that are defined by a shaded line (PR and STR) indicate length and sequence variation in these regions. Diamonds indicate potential glycosylation sites; (†) catalytic active site residues D_{214}, H_{252}, S_{423}; (|) conserved cysteine residues. Residues were numbered with reference to the PRT1(73j) sequence.

Figure 7

Recombinant PRT1 polypeptides, expressed in *E. coli* as thioredoxin fusion proteins, separated by SDS-PAGE and cross-reacted with an anti-thioredoxin antibody. *E. coli* DE3(BL21) transformed with: lane 1: control plasmid pET32a; lane 2: F1a1a (portion of pro-domain of *P.carinii* sp. f. *carinii PRT1* gene); lane 3: G1b1c (portion of P-domain of *P.carinii* sp. f. *carinii PRT1* gene); lane 4: H1a1a (portion of catalytic domain of *P.carinii* sp. f. *hominis PRT1* gene).

CLAIMS

1. An isolated DNA comprising part or all of a *PRT1* gene of a non-rat infecting species of *Pneumocystis carinii.*

2. The DNA according to claim 1, comprising part or all of a *PRT1* gene of a human-infecting species of *Pneumocystis carinii.*

3. The DNA according to claim 1 or claim 2, wherein the *PRT1* gene is in the form of cDNA.

4. An isolated DNA comprising a sequence shown in figure 1, or a non-rat sequence shown in figure 2, or a sequence which hybridises to either of these under stringent conditions.

5. The DNA according to claim 1 or claim 4, wherein the *PRT1* gene has been mutated by point mutation, deletion, insertion, or other means.

6. A recombinant vector containing the DNA according to any one of claims 1 to 5.

7. A recombinant polypeptide which is part or all of a *PRT1* gene product, expressed by a vector according to claim 6.

8. Synthetic peptides corresponding to antigenic portions of a PRT1 gene product.

9. A synthetic peptide chosen from:

TWRDVQALIVETAVP	(SEQ ID NO: 16)
ITSPSGVTSVLAHRR	(SEQ ID NO: 17)
ESEGVPPPSYPFLSR	(SEQ ID NO: 18)
ASTPLAAGVIALLLS	(SEQ ID NO: 19)
FRGESIVGNWTIDVE	(SEQ ID NO: 20)
DNQHIFSIEKGVLED	(SEQ ID NO: 21)

10. A method of producing antibodies specifically immunoreactive with a *Pneumocystis carinii* protease, which method comprises using a polypeptide according to claim 7 or a synthetic peptide according to claim 8 or claim 9 to generate an immune response.

11. Antibodies produced by the method according to claim 10.

12. Antibodies according to claim 11, which are monoclonal.

13. A method of screening for anti-*Pneumocystis carinii* compounds, which method comprises providing a source of a recombinant polypeptide expressed by part or all of a *PRT1* gene or cDNA, and contacting the compound with the recombinant polypeptide.

14. The method according to claim 13, wherein the recombinant polypeptide is expressed at the surface of a cell.

15. The method according to claim 13 or claim 14, for screening for protease inhibitors effective against *Pneumocystis carinii.*

16. The method according to any one of claims 13 to 15, using a recombinant polypeptide corresponding to part or all of the catalytic domain of the protease.

17. A cell transfected with a vector according to claim 6 and expressing a polypeptide according to claim 7.

18. An engineered cell line expressing a recombinant polypeptide from part or all of a *PRT1* gene or cDNA, which may be mutated by point mutation, deletion, insertion or other means, useful in the method according to any one of claims 13 to 16.

19. The cell line according to claim 18, wherein the *PRT1* gene or cDNA is from a human-infecting *Pneumocystis carinii* species.

20. The method according to any one of claims 13 to 16, wherein the *PRT1* gene or cDNA has been mutated by point mutation, deletion, insertion or other means.

21. A *Pneumocystis carinii* protease isolated using an antibody according to claim 11 or claim 12.

22. A *PRT1* clone for part or all of the human-infecting *Pneumocystis carinii PRT1* gene.

Figure 1

Human-derived *Pneumocystis carinii* subtilisin-like serine protease (*PRT1*) (H13)

```
1    TGAAGTAGCT GCCGTTCGAA ATACTGTTTG TGGAATCGGT GTTGCATATG
51   AATCCAAAGT TTCTGGTATT TTATTCTTTT TGACTGAATC TAATATAATA
101  TCATTAAGGT TTGCGAATAT TATCCGGGCC TATAACAGAT CTTGATGAAG
151  CAGAATCGCT TAATTATGAT TTCCATAAAA ATCATATTTA TTCCTGTAGT
201  TGGGGACCTG ACGATGATGG AAAAACTGTT GATGGGCCTT CTTCTCTTGT
251  TCTTAGAGCA CTTATTAATG GAGTAAATAA TGGAAGGAAT GGGTTGGGTT
301  CTATCTATGT TTTTGCATCA GGAAATGGTG GAATATATGA AGATAACTGT
351  AATTTCGATG GATATGCAAA TAGTGTGTTT ACCATTACTA TTGGTGGCAT
401  AGATAAACAT GGAAAGCGTC TTAAATATTC TGAAGCGTGT TCTTCTCAGC
451  TAGCTGTTAC ATATGCAGGT GGAAGTGCGG ATATATTTGT AACTTTAATT
501  CTATTTTTTT TTATATAAAT TTATAATAAT TAGTATACTA CTGATGTTGG
551  TACAAATAAA TGTACGAGTA GACATGGTGG TACC
```

Figure 2

```
  1-3paga  A G A A G T G G C A G G C G C C A G G A A T G A T T T T T G T G G G C T T G G T G T T G C A T A T G   50
  1-3-73j  A G A A G T G G C A G C C G C C A G G A A T G A T T T T T G T G G G C T T G G T G T T G C A T A T G   50
1-3prp5e  A G A A G T G G C A G G C G C C A G G A A T G A T T T T T G T G G G C T T G G T G T C G C A T A T G   50
 rv5pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
rv16pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
 m1pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
 m7pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
m13pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
  hpcprt1  T G A A G T A G C T G C C G T T C G A A A T A C T G T T T G T G G A A T C G G T G T T G C A T A T G   50

  1-3paga  A A T C T A A T A T T T C A G G T A T T T T - - - - T C T T T A A T T G G T A C C T A T C T A A T A   96
  1-3-73j  A A T C T A A T A T T T C A G - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   65
1-3prp5e  A A T C T A A T A T T T C A G - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   65
 rv5pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
rv16pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
 m1pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
 m7pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
m13pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
  hpcprt1  A A T C C A A A C T T T C T G G T A T T T T A T T C T T T T T G A C T G A A T C T A A T A T A A T A  100

  1-3paga  T T G T T A A G G A T T A C G A T T T A T G C C T T C T G C T C G T T C G T C T T G G C T T G A A G  146
  1-3-73j  - - - - - - - - G A T T A C G A T T T T T G C C T T C T G G T C T C T C G T A T C A T C T T G A G T  107
1-3prp5e  - - - - - - - - G A T T A C G A T T T A T G C C T T C T G C T C G T T C G T C T T G G C T T G A A G  107
 rv5pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
rv16pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
 m1pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
 m7pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
m13pcprt1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
  hpcprt1  T C A T T A A G G T T T G C G A A T A T T A T C C G G G C C T A T A A C A G A T C T T G A T G A A G  150
```

Figure 2

```
   1-3paga G A G A A G C T C T T A T T T A C A A A T A T G A T G T T A A T C A T A T T T A T T C T T G T A G C   196
   1-3-73j C A C T A G C T C T T A G T T A T A A A C C G A A T G T T A A T T A T A T T T A T T C T T G T A G C   157
  1-3prp5e G A G A A G C T C T T A T T T A C A A A T A C G A T G T T A A T C A T A T T T A T T C T T G T A G C   157
 rv5pcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -     0
rv16pcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -     0
  m1pcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -     0
  m7pcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -     0
 m13pcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -     0
   hpcprt1 C A G A A T C G C T T A A T T A T G A T T T C C A T A A A A A T C A T A T T T A T T C C T G T A G T   200

   1-3paga T G G G G A C C T G C C G A T A C T G G G A A T T T A A C T C A A G A T A T T T T T T A T A C T A C   246
   1-3-73j T G G G G A C C T C C T G G T G A T G G A T A T G C A G C T A T C C C A A T G T A T C C T A C T A C   207
  1-3prp5e T G G G G A C C C G C C G A T A C T G G G A A T T T A A C T C A A G A T A T T T T T T A T A C T A C   207
 rv5pcprt1 - - - - - - - C T C C T G A T A G T G G A A G A T A T G C T G A A G A C A T T A A T G T T C C T A C    43
rv16pcprt1 - - - - - - - C T G C T G A T T T T G G A G T T A C A G C T G A G G A C A T T T A C A C T C C T A C    43
  m1pcprt1 - - - - - - - C T C G T G A T G A T G G A A A A A C A A T T G A A G G A G T T C C T T A T A G T G C    43
  m7pcprt1 - - - - - - - C T C G T G A T G A T G G A A A A A C A A T T G A A G G A G T T C C T T A T A G T G C    43
 m13pcprt1 - - - - - - - C T C G T G A T G A T G G A A A A A C A A T T G A A G G A G T T C C T T A T A G T G C    43
   hpcprt1 T G G G G A C C T G A C G A T G A T G G A A A A A C T G T T G A T G G G C C T T C T T C T C T T G T   250

   1-3paga T T A T T C T G C A A T T A T T A A A G G G A T A A A T C A A G G A A G G A A T G G T C T T G G T T   296
   1-3-73j T T A T T C T G C A A T T A T T A A A G G G A T A A A A G A A G G A A G G A A C G G T C T T G G C T   257
  1-3prp5e T T A T T C T G C A A T T A T T G A A G G G A T A A A T C A A G G A A G G A A T G G T C T T G G T T   257
 rv5pcprt1 T T A T A A A G C A A T T A T T A A A G G G G T A A A T G A A G G A A G G A A T G G T C G T G G T G    93
rv16pcprt1 T T A T A A A G T A A T T C T T A A A G G G G T A A C T G A A G G A A G G A A C G G T C T T G G T T    93
  m1pcprt1 A T A T A A T T C A A T T A T T A A T G G G A T A A A T C T T G G A A G G A A A G G T C T T G G T T    93
  m7pcprt1 A T A T A A T T C A A T T A T T A A T G G G A T A A A T C T T G G A A G G A A A G G T C T T G G T T    93
 m13pcprt1 A T A T A A T T C A A T T A T T A A T G G G A T A A A T C T T G G A A G G A A A G G T C T T G G T T    93
   hpcprt1 T C T T A G A G C A C T T A T T A A T G G A G T A A A T A A T G G A A G G A A T G G G T T G G G T T   300
```

Figure 2

```
   1-3paga C T A T A T A C G T T T T C G G G T C A G G A A A T G G T G G - - - A T A T T T T G A T A A T T G T   343
   1-3-73j C T A T A T A T G T T T T T G G A A C C G G A A A T G G T G G - - - A T C A T T G G A T G G T T G T   304
  1-3prp5e C T A T A T A C G T T T T C G G G T C A G G A A A T G G T G G - - - A T A T T T T G A T A A T T G T   304
 rv5pcprt1 C T T T A T A T G T T T T T G G G T C T G G A A A T G G T G G - - - C T T T A T T G A T G A T T G T   140
rv16pcprt1 C T A T A T A T G T T T T T G G G T C T G G A A A T G G T G G - - - C A A A T T T G A T G A T T G T   140
  m1pcprt1 C T A T A T A T G T T T T T G G A A G T G G A A A T G G A G G - - - C T A T T A T G A T A A T T G C   140
  m7pcprt1 C T A T A T A T G T T T T T G G A A G T G G A A A T G G A G G - - - C T A T T A T G A T A A T T G C   140
 m13pcprt1 C T A T A T A T G T T T T T G G A G G T G G A A A T G G A G G - - - C T A T T A T G A T A A T T G C   140
   hpcprt1 C T A T C T A T G T T T T T G C A T C A G G A A A T G G T G G A A T A T A T G A A G A T A A C T G T   350

   1-3paga A A T T A C G A T G G A T A T G C A A A T A G C C C A T A T A C T A T T A C T A T C G C T G C T A T   393
   1-3-73j A A T T A C G A T G G A T A T G C A A A T A G T C C A T A T A C T A T T A C T A T C G C T G C T A T   354
  1-3prp5e A A T T A C G A T G G A T A T G C A A A T A G C C C A T A T A C T A T T A C T A T C G C T G C T A T   354
 rv5pcprt1 A A T T A C G A T T C A T T T G C A A G T A G T C C A T A T A C T A T T A C T A T T G C T G C T A T   190
rv16pcprt1 A A T T A C G A T G G T T A T G C A A G T A G T C C A T A T A C T G T T A C T G T T T C T T C T A T   190
  m1pcprt1 A A T T A C G A T G G A T A T G T A G T T A G T C C A T A T A C T A T T A C T A T C G G T T C T A C   190
  m7pcprt1 A A C T A C G A T G G A T A T G T A G T T A G T C C A T A T A C T A T T A C T A T C G G T T C T A T   190
 m13pcprt1 A A T T A C G A T G G A T A T G T A G T T A G T C C A T A T A C T A T T A C T A T C G G T T C T A T   190
   hpcprt1 A A T T T C G A T G G A T A T G C A A A T A G T G T G T T T A C C A T T A C T A T T G G T G G C A T   400

   1-3paga A G A T G C A G A A G A A A A A A G A T T C A T A T T T T C A G A G C C A T G T C C T T G T A T T T   443
   1-3-73j A G A T T C A G A A G A T A A A A A T T T T T A T T T T T C A G A G T C A T G T C C T T G C A T T T   404
  1-3prp5e A G A T G C A G A A G A A A A A A G A T T C A T A T T T T C A G G G C C A T G T C C T T G T A T T T   404
 rv5pcprt1 C A A T G C A A G T G A T G C A A A A T T T A G A T A T T C A G A G T C A T G T T C T G C T A T C T   240
rv16pcprt1 C A A T G C A G A T G A T A A T A A T T T T G A T T T T A T A G A G C C A T G T T C T G C T A T C T   240
  m1pcprt1 A G A T G T G A G A G G A A T A A G A C A T T A T T T T T C A G A G C A A T G T T C T T C C G T T C   240
  m7pcprt1 A G A T G T G A G A G G A A T A A G A C A T T A T T T T T C A G A G C A A T G T T C T T C C G T T C   240
 m13pcprt1 A G A T G T G A G A G G A A T A A G A C A T T A T T T T T C A G A G C A A T G T T C T T C C G T T C   240
   hpcprt1 A G A T A A A C A T G G A A A G C G T C T T A A A T A T T C T G A A G C G T G T T C T T C T C A G C   450
```

Figure 2

```
1-3paga     TAGCTTCTACGTATTCTGGCAAGCGTG------GTGCATATATTGTAATC   487
1-3-73j     TGGCTTCTACATATTCTGGC------GGAGAAAATGGATCTATT------   442
1-3prp5e    TAGCTTCTACGTATTCTGGCAAGCGTG------GTGCATATATT------   442
rv5pcprt1   TGGCTTCTACATATTCTGGCGCTGG------GAAATCGCCTATTGTAACC   284
rv16pcprt1  TGGCTTCTACATATTCTGGCGCTGG------GAAATGGCCTATTGTAACC   284
m1pcprt1    TTGCTTCTACATATTCGGGTTCTATTGTAACCAATGCACGCATTGTAATT   290
m7pcprt1    TTGCTTCTACATATTCGGGTTCTATTGTAACCAATGCACGCATTGTAATT   290
m13pcprt1   TTGCTTCTACATATTCGGGTTCTATTGTAACCAATGCACGCATTGTAATT   290
hpcprt1     TAGCTGTTACATATGCAGGT------GGAAGTGCGGATATATTTGTAACT   494

1-3paga     ----TTTTCTTTTTTTTTATAATAAATTGATCGT-TTTAGTATACTACGG   532
1-3-73j     ---------------------------------------TATACTACGGA   453
1-3prp5e    ---------------------------------------TATACTACGGA   453
rv5pcprt1   ----TTTTCTTATTTTTTATAATAAATTTATTTT-TTTAGTATACTACGG   329
rv16pcprt1  ----TTTTCTTATTTTTTATAATAAATTTATTTT-TTTAGTATACTCCAG   329
m1pcprt1    ----TTTCCTTTTTTTTTATAATAAATTTATCTT-GTTAGTATACTACGG   335
m7pcprt1    ----TTTCCTTTTTTTTTATAATAAATTTATCTT-GTTAGTATACTACGG   335
m13pcprt1   ----TTTCCTTTTTTTTTATAATAAATTTATCTT-GTTAGTATACTACGG   335
hpcprt1     TTAATTCTATTTTTTTTTAT-ATAAATTTATAATAATTAGTATACTACTG   543

1-3paga     ATGTTGGTACGACAGAATGCAGCATTAGACATACTGGAAGTTCTGCTTCT   582
1-3-73j     TCTTGGTAAGGAGGGATGCACTACTGAACATACTGGAGCTTCTGCTTCTA   503
1-3prp5e    TGTTGGTACGACAAAATGCAGCATTAGACATACTGGAAGTTCTGCTTCTA   503
rv5pcprt1   ACGT---TGAAAATAAATGTACCACTATGCATACTGGAACGTCTGCTTCT   376
rv16pcprt1  ACGTGGGGGAATCTAGGTGTTCCACTAAGCATACTGGATCGTCTGCTTCT   379
m1pcprt1    ATGTTGGTGAGAAAGGATGCAGCACTGTACATTCTGGATCTTCTGCTTCC   385
m7pcprt1    ATGTTGGTGAGAAAGGATGCAGCACTGTACATTCTGGATCTTCTGCTTCC   385
m13pcprt1   ATGTTGGTGAGAAAGGATGCAGCACTGTACATTCTGGATCTTCTGCTTCC   385
hpcprt1     ATGTTGGTACAAATAAATGTACGAGTAGACATGGTGGTACC---------   584
```

Figure 2

```
   1-3paga A C A C C T C T T G C T G C G G G T G T T A T T G C T C T T C T T C T T T C A G C A T G G T A A G A   632
   1-3-73j C A C C T C T T G C T G C G G G T A T T A T T G C T C T T G T T C T T T C A G C G A A - - - - - - -   546
  1-3prp5e C A C C T C T T G C T G C G G G T G T T A T T G C T C T T C T T C T T T C A G C A T G - - - - - - -   546
 rv5pcprt1 A C A C C T A T T G C T T C A G G T A T C A T G G C C C T T C T T A T A T C A G C A T G G T A A G A   426
rv16pcprt1 G T A C C T A T T G C T G C A G G T A T C A T A G C C C T T G C T C T A T C A G T A C G G T A A G A   429
  m1pcprt1 A C A C C T A T T G C A G C A G G C G T T A T A G C T C T T G T T C T T T C A G T A C G G T A A G A   435
  m7pcprt1 A C A C C T A T T G C A G C A G G C G T T A T A G C T C T T G T T C T T T C A G T A C G G T A A G A   435
 m13pcprt1 A C A C C T A T T G C A G C A G G C G T T A T A G C T C T T G T T C T T T C A G T A C G G T A A G A   435
   hpcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   584

   1-3paga A T A T C A T T A A A A T T A T T T G A C - T A A A A A A T T A G                                     664
   1-3-73j - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -                                     546
  1-3prp5e - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -                                     546
 rv5pcprt1 A T T T A A T T A A T A T T A T T C A A C C T A A A T A T T T A G                                     459
rv16pcprt1 A T T T A A T T A A T A T T A T T C A A C C T A A A T A T T T A G                                     462
  m1pcprt1 A T T C C - - - - - - - - - - - - - - - - - - - - - - - - - - - -                                     440
  m7pcprt1 A T T C C - - - - - - - - - - - - - - - - - - - - - - - - - - - -                                     440
 m13pcprt1 A T T C C - - - - - - - - - - - - - - - - - - - - - - - - - - - -                                     440
   hpcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -                                     584
```

Figure 3

```
  pagaprt1 E V A G A R N D F C G L G V A Y E S N I S G L R F M P S A R S S W L E G E A L I Y K Y D V N H I Y S   50
    73jprt1 E V A A A R N D F C G L G V A Y E S N I S G L R F L P S G L S Y H L E S L A L S Y K P N V N Y I Y S   50
 prp5eprt1 E V A G A R N D F C G L G V A Y E S N I S G L R F M P S A R S S W L E G E A L I Y K Y D V N H I Y S   50
 rv5pcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
rv16pcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
 m1pcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
  m7pcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
m13pcprt1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
   hpcprt1 E V A A V R N T V C G I G V A Y E S K V S G L R I L S G P I T D L D E A E S L N Y D F H K N H I Y S   50

  pagaprt1 C S W G P A D T G N L T Q D I F Y T T Y S A I I K G I N Q G R N G L G S I Y V F G S G N G G Y F - D   99
    73jprt1 C S W G P P G D G Y A A I P M Y P T T Y S A I I K G I K E G R N G L G S I Y V F G T G N G G S L - D   99
 prp5eprt1 C S W G P A D T G N L T Q D I F Y T T Y S A I I E G I N Q G R N G L G S I Y V F G S G N G G Y F - D   99
 rv5pcprt1 - - - - - P D S G R Y A E D I N V P T Y K A I I K G V N E G R N G R G A L Y V F G S G N G G F I - D   44
rv16pcprt1 - - - - - A D F G V T A E D I Y T P T Y K V I L K G V T E G R N G L G S I Y V F G S G N G G K F - D   44
 m1pcprt1 - - - - - R D D G K T I E G V P Y S A Y N S I I N G I N L G R K G L G S I Y V F G S G N G G Y Y - D   44
  m7pcprt1 - - - - - R D D G K T I E G V P Y S A Y N S I I N G I N L G R K G L G S I Y V F G S G N G G Y Y - D   44
m13pcprt1 - - - - - R D D G K T I E G V P Y S A Y N S I I N G I N L G R K G L G S I Y V F G G G N G G Y Y - D   44
   hpcprt1 C S W G P D D D G K T V D G P S S L V L R A L I N G V N N G R N G L G S I Y V F A S G N G G I Y E D  100

  pagaprt1 N C N Y D G Y A N S P Y T I T I A A I D A E E K R F I F S E P C P C I L A S T Y S G K R G A Y I Y T  149
    73jprt1 G C N Y D G Y A N S P Y T I T I A A I D S E D K N F Y F S E S C P C I L A S T Y S G G E N G S I Y T  149
 prp5eprt1 N C N Y D G Y A N S P Y T I T I A A I D A E E K R F I F S G P C P C I L A S T Y S G K R G A Y I Y T  149
 rv5pcprt1 D C N Y D S F A S S P Y T I T I A A I N A S D A K F R Y S E S C S A I L A S T Y S G A G K S P I Y T   94
rv16pcprt1 D C N Y D G Y A S S P Y T V T V S S I N A D D N N F D F I E P C S A I L A S T Y S G A G K W P I Y T   94
 m1pcprt1 N C N Y D G Y V V S P Y T I T I G S T D V R G I R H Y F S E Q C S S V L A S T Y S G S - - - - I Y T   90
  m7pcprt1 N C N Y D G Y V V S P Y T I T I G S I D V R G I R H Y F S E Q C S S V L A S T Y S G S - - - - I Y T   90
m13pcprt1 N C N Y D G Y V V S P Y T I T I G S T D V R G I R H Y F S E Q C S S V L A S T Y S G S - - - - I Y T   90
   hpcprt1 N C N F D G Y A N S V F T I T I G G I D K H G K R L K Y S E A C S S Q L A V T Y A G G S A D I F Y T  150
```

Figure 3

```
pagaprt1   T D V G T T E C S I R H T G S S A S T P L A A G V I A L L L S A   181
73jprt1    T D L G K E G C T T E H T G A S A S T P L A A G I I A L V L S A   181
prp5eprt1  T D V G T T K C S I R H T G S S A S T P L A A G V I A L L L S A   181
rv5pcprt1  T D V - E N K C T T N H T G T S A S T P I A S G I M A L L I S A   125
rv16pcprt1 P D V G E S R C S T K H T G S S A S V P I A A G I I A L A L S -   125
m1pcprt1   T D V G E K G C S T V H S G S S A S T P I A A G V I A L V L S -   121
m7pcprt1   T D V G E K G C S T V H S G S S A S T P I A A G V I A L V L S -   121
m13pcprt1  T D V G E K G C S T V H S G S S A S T P I A A G V I A L V L S -   121
hpcprt1    T D V G T N K C T S R - - - - - - - - - - - - - - - - - - - - -   161
```

Fig.4 (Cont i).

```
Paga      G S S A S T P L A A G V I A L L L S A C P N L T W R D I Q A L I V E T A V P F N P S H P D W D D L P S G R R Y N N F F G   471
73j       G A S A S T P L A A G I I A L V L S A N P N L T W H D V Q A L I V E T A V P F N L E Y P G W D K L P S E R H Y S N N F G   480
Prp5e     G S S A S T P L A A G V I A L L L S A C P N L T W R D I Q A L I V E T A V P F N P S H P D W D D L P S G R R Y N N F F G   473
N14       - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -     0
Prp1a     G T S A S V P I A A G I I A L V L S V N P N L T W - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   132
Prp3a     G I S A A I P I A S A I I A L V L S V N P N L T - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   132
Prp7a     G T P A S T P I A A G I I A L V L S V N P N L T - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   131
Prp2c     - - - - - - - - - - G V I A L H L S A C P N L T W R D I Q A L I V E T A V P F N P S H P D W D D L P S G R R Y N N F F G    50
Prp3c     - - - - - - - - - - G V I A L L L S A C P N L T W R D I Q A L I V E T A V P F N P S K P D W D D L P S G R R Y N N F F G    50
Prp4c     - - - - - - - - - - G V I A L L L S A C P N L T W R D I Q A L I V E T A V P F N P S K P D W D D F P S G R R Y N N F F G    50

Paga      Y G K L D A Y R H V E K A R T F K T L N P Q T H F S T Q L I P L N K K F S E N G G H I T S S F Y I H R G Y P K H Y K F K   531
73j       F G K L D A Y R H V E R A K T F K T L N A Q T H F S T Q L I P L N K T F S E N G G H I T S T F Y I D S G S P T H Y N F K   540
Prp5e     Y G K L D A Y R H V E K A R T F K T L N P Q T H F S T Q L I P L N K K F S E N G G H I T S S F Y I H R G Y P K H - - - -   529
N14       - - - - - - Y R H V E K A K T F K T L N A Q T H F S T Q L I Q I N H K F P D F S R R I T S S F Y I H S G Y P T H Y N F K    54
Prp1a     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   132
Prp3a     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   132
Prp7a     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   131
Prp2c     Y G K L D A Y R H D E K A R T F K T L N P Q T H F S T Q L I P L N K K F S E N G G H I T S S F Y I H R G Y P K - - - - -   105
Prp3c     Y G K L D A Y R H V E K A R T F K T L N P Q T H F S T Q L I P L N K K F S E N G G H I T S S F Y I H R G Y P K - - - - -   105
Prp4c     Y G K L D A Y R H V E K A R T F K T L N P Q T H F S T Q L I P L N K K F S E N G G H I T S S F Y I H R G Y P K - - - - -   105
Prptaf2   - - - - - A Y R H V E K A R T F K T L N P Q T H F S T Q L I P L N K K F S E N G G H I T S S F Y I H R G Y P K H Y K F K    55
Prpf4     - - - - - A Y R H V E K A R A F K T L N P Q T H F S T Q L I P L N K K F S E N G G H I T S S F Y I H C G Y P K H Y K F K    55
Prp5f     - - - - - A Y R H V E K A R T F K T L N P Q T H F S T Q L I P L N K K F S E N G G H I T S S F Y I H R G Y P K H Y K F K    55

Paga      S L E Y V G V S F H Y Q H Q R R G H L E F N I T S P S G V T S V L A H R R N R D K H G G S I L W T F H T V K H W G E S I   591
73j       N L E Y V G V S F H Y K H Q Y K G H L E F N I T S P S G V T S V L A H R R T N D Y N S G T F H W F F T T Y K H W G E T I   600
Prp5e     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   529
N14       N L E C V G V S F H Y Q H Q K R G R L E F S I T S P A N V T S K L A R V K V R D E E S G T F S W I F T T Y K H W G E K I   114
Prp2c     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   105
Prp3c     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   105
Prp4c     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   105
Prptaf2   S L E Y V G V S F H Y Q H Q R R G H L E F N I T S P S G V T S V L A H R R N R D K H G G S I L W T F H T V K H W G E S I   115
Prpf4     S L E Y V G V S S H Y Q H Q R R G H L E F N I T S P S G V T S V L A H R R N R D K H G G S I L W T F H T V K H W G E S I   115
Prp5f     S L E Y V G V S F H Y Q H Q R R G H L E F N I T S P S G V T S V L A H R R N R D K H G G S I L W T F H T V K H W G E S I   115

Paga      V G N W T I D V E D K K D E N L D G G V F D W Q L H F F G E S C E S E G V - P P P S Y P F L S R Y - - - - - - - - P T T   642
73j       V G N W T I D V E D E K V S N L D G E I F D W Q L H F F G E S T D S S K A E L H P P Y P F K - - - - - - - - - - - - - -   646
Prp5e     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   529
N14       V G N W T I D V E D E K D P N L D G E V F N W Q L H F F G E S I D S T K A - Q P P P Y P F V H K Q P T T K P P P E P T T   173
Prp2c     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   105
Prp3c     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   105
Prp4c     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   105
Prptaf2   V G N W T I D V E D K K D E N L D G G V F D W Q L H F F G E S C E S D R Y R L L H I L F Y L D I Q L L R L H Q I Q H L H   175
Prpf4     V G N W T I D V E D K K D E N L D G G V F D W Q L H F F G E S C E S E R Y R L L H I L F Y L D I Q L L R L H Q I Q H L H   175
Prp5f     V G N W T I D V E D K K D E N L D G G V F D W Q L H F F G E S C E S E R Y R L L H I L F Y L D I Q L L R L H Q I Q H L H   175
```

Fig.4.

```
 Paga  M I F K I L I T F F L Y W I Y L V R V R C E K K P V D F E N N D Y Y - H F H F S E D V D I E E F S R A V G L K Y H N K V   59
  73j  M I F K I L L T F F L Y W I Y L V R V R C E K V P V D F E N N D Y Y Y Y F H L S E D V D I E E F S R A V G F K Y H N K V   60
Prp5e  M I F K I L I T F F L Y W I Y L V R V R C E K V P I D F E N N D Y Y - H F H F S G D V D I E D F S R A L G F K H Y N K L   59

 Paga  E Y L D N Q H I F F I E K G V L E D E I K E K I E N Y F G L E K G R N A I D G F N S D K L F Y Y E K Q K L V K R V N R G  119
  73j  D H L D N H H I F F I E K G V L E D E I K E K I E N Y F S L E K G R N A I D G F N S D K L F Y Y E K Q K L V K P V N R G  120
Prp5e  E H L D N Q H I F S I E K G V L E D E I K E K I E N Y F G L E K G R N A I N G F N S D K L F Y Y E K Q K L V K R E N R G  119

 Paga  V I R D D I Y F D N E G L Y N R R I V K N - V V K D S T G D Q A - - - - V D L R E K I K K I K E E L N I S D F Y F D K Q  174
  73j  A I R D D I Y F D N Q D L Y N D E E I V N N V V K D P T V D Q A K K S T E D L K E R L K K I K K E L G I S D P C F D K Q  180
Prp5e  V I R D D I Y F D N K G L Y N R R V V K N - V V K D P T V D - - - - L P V N L T Q K L K K I K E E L N I S N P Y F D K Q  174

 Paga  W - - F N K D K A G V D I N V T G V W L Q G I K G K N V T V A I V D D G L D Y T N K D L A P H Y N A N A S Y N F A S K T  232
  73j  W Y L F N T E K P G V D I N V T G V W L Q G I T G K G V T V A I A D N G L D Y T N K D L A P H Y N S Q G S Y D F V S K T  240
Prp5e  W Y L F N K D K A G V D I N V T G V W L Q G I K G K N V T V A I V D D G L D Y T N K D L A P H Y N A N A S Y N F A S K T  234

 Paga  G D P K P E P - S D T H G T K C A G E V A G A R N D F C G L G V A Y E S N I S G L R F K P S A R S S W L E G E A L I Y K  291
  73j  D D P N P K S S S D T H G T R C A G E V A A A R N D F C G L G V A Y E S N I S G L R F L P S G L S Y H L E S L A L S Y K  300
Prp5e  G D P K P G P - S D T H G T K C A G E V A G A R N D F C G L G V A Y E S N I S G L R F H P S A R S S W L E G E A L I Y K  293

 Paga  Y D V N H I Y S C S W G P A D T G N L T Q D I F Y T T Y S A I I K G I N Q G R N G L G S I Y V F G S G N G G Y F D N C N  351
  73j  P N V N Y I Y S C S W G P P G D G Y A A I P H Y P T T Y S A I I K G I K E G R N G L G S I Y V F G T G N G G S L D G C N  360
Prp5e  Y D V N H I Y S C S W G P A D T G N L T Q D I F Y T T Y S A I I K G I N Q G R N G L G S I Y V F G S G N G G Y F D N C N  353
  M14  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
Prp1a  - - - - - - - - - - - - - S D S G N T T G P I P Y L S Y S S I I E K L N Q G R K G L G S I Y V F G S G N G G P S D D C N   47
Prp3a  - - - - - - - - - - - - L L D L G N K I E N I F Y S T Y T A I I D G I N K G R K G L G S I Y V F G S G N G G D L D H C N   48
Prp7a  - - - - - - - - - - - - - P E Y G F L A N P M N P T T Y S A I I K G V K E G R N G L G S I Y V F G T G N G G L L N D C N   47

 Paga  Y D G Y A N S P Y T I T I A A I D A E E K R F I F S E P C P C I L A S T Y S G K R G A Y I Y T T D V G T T E C S I K H T  411
  73j  Y D G Y A N S P Y T I T I A A I D S E D K N F Y F S E S C P C I L A S T Y S G G E N G S I Y T T D L G K E G C T T E H T  420
Prp5e  Y D G Y A N S P Y T I T I A A I D A E E K R F I F S G P C P C I L A S T Y S G K R G A Y I Y T T D V G T T K C S I K H T  413
  M14  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -    0
Prp1a  Y D G Y A N S P Y T I T I G A I D A D E K I F Y F S E P C P C I L A S T Y S G K R G A Y I H T T D V G L A G C T F K H S  107
Prp3a  Y D G Y A N S P Y T I T I A A I N A R E R K F S F S E P C P C I L A S T Y S G G E S G R I Y T T D V G S T A C T S I H T  108
Prp7a  Y D G Y A N S P Y T I T I A A I N S E D E R P Y F S E S C P C I L A S T Y S G G G N G S I Y T T D V G I T N C T A K H T  107
```

Fig.4 (Cont ii).

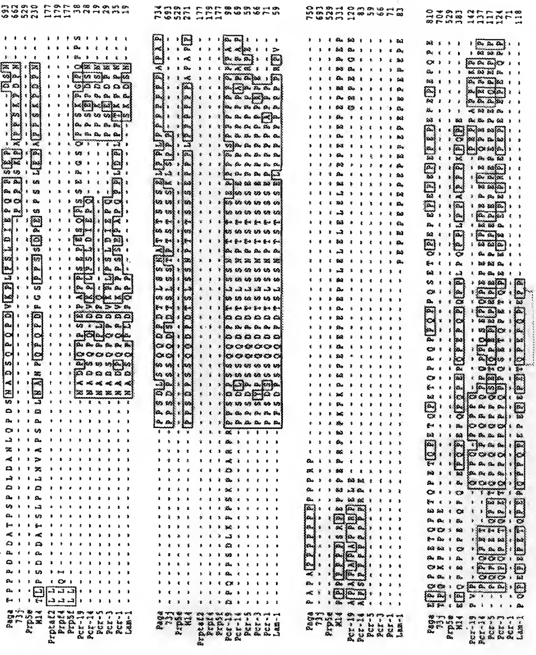


Fig.4 (Cont iii).

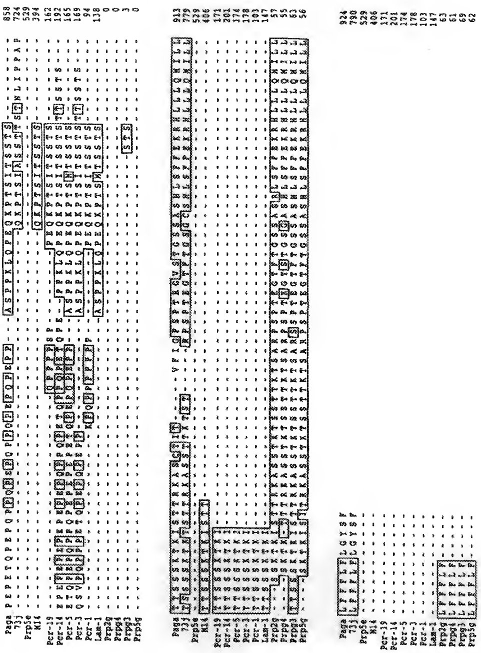


Figure 5

```
 Name: Paga              Len:  3150  Check: 9848  Weight:  1.00
 Name: 73j               Len:  3150  Check: 2744  Weight:  1.00
 Name: Prp5e             Len:  3150  Check: 2286  Weight:  1.00
 Name: M14               Len:  3150  Check: 9011  Weight:  1.00
 Name: Prp2g             Len:  3150  Check: 9244  Weight:  1.00

//

        1                                                   50
  Paga  ATGATTTTTA AGATACTCAT TACTTTTTTC TTATACTGGA TCTATTTAGT
   73j  ATGATTTTCA AGATACTCCT TACTTTTTTC TTATACTGGA TCTATTTAGT
 Prp5e  ATGATTTTTA AGATACTCAT TACTTTTTTC TTATACTGGA TCTATTTAGT
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        51                                                 100
  Paga  TAGAGTAAGA TGTGAAATGA AGCCAGTAGA CTTTGAAAAT AATGATTATT
   73j  TAGAGTAAGA TGTGAAATGG TGCCAGTAGA CTTTGAGAAT AATGATTATT
 Prp5e  TAGAGTAAGA TGTGAAATGG TGCCAATAGA CTTTGAGAAT AATGATTATT
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        101                                                150
  Paga  A...TCATTT TCATTTCTCA GAAGATGTTG ATATTGAGGA GTTTTCGCGG
   73j  ATTATTATTT TCATCTCTCA GAAGATGTTG ATATTGAGGA GTTTTCTCGG
 Prp5e  A...TCATTT TCATTTCTCA GGAGATGTTG ATATTGAGGA TTTTTCGAGG
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        151                                                200
  Paga  GCGGTAGGAT TGAAATATCA TATGAAAGTA GAATATCTGG ATAACCAGCA
   73j  GCGGTAGGAT TCAAATATCA TATGAAAGTA GATCATCTGG ATAACCACCA
 Prp5e  GCGTTAGGAT TTAAACATTA TATGAAACTA GAACATCTGG ATAACCAGCA
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        201                                                250
  Paga  TATATTTTTC ATAGAAAaGG GTGTTTTAGA AGACGAAATT AAAGAAAAAA
   73j  TATATTTTTT ATAGAAAAGG GTGTTTTAGA AGACGAAATT AAAGAAAAAA
 Prp5e  TATATTTTCT ATAGAAAAGG GTGTTTTAGA AGACGAAATT AAAGAAAAAA
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        251                                                300
  Paga  TTGAGAATTA TTTTGGTTTA GAAAaAGGAA GAAaTGCAaT AGATGGGTTT
   73j  TTGAGAATTA TTTCAGTTTA GAAAAAGGAA GAAATGCAAT AGATGGGTTT
 Prp5e  TTGAGAATTA TTTTGGTTTA GAAAAAGGAA GAAATGCAAT AAATGGGTTT
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        301                                                350
  Paga  AATAGTGACA AACTTTTTTA TTATGAGAAA CAAAAGTTGG TCAAGCGAGT
   73j  AATAGTGACA AGCTTTTTTA TTATGAGAAA CAAAAGTTGG TCAAGCCAGT
 Prp5e  AATAGTGACA AGCTTTTTTA TTATGAGAAA CAAAAGTTGG TCAAGCGAGA
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        351                                                400
  Paga  AAACAGGGGT GTGATAAGAG ACGATATATA TTTTGATAAT GAAGGTCTTT
   73j  AAACAGGGGT GCGATAAGAG ACGATATATA TTTTGATAAC CAAGATCTTT
 Prp5e  AAACAGGGGT GTGATAAGAG ACGATATATA TTTTGATAAT AAAGGTCTTT
   M14  .......... .......... .......... .......... ..........
```

Figure 5

```
Prp2g   .......... .......... .......... .......... ..........

        401                                                450
 Paga   ATAATAGAAG AA...TTGTT AAGAATGTTG TAAAAGATTC GACGGGAGAT
  73j   ATAATGATGA AGAAATTGTC AATAATGTTG TAAAAGATCC GACGGTAGAT
Prp5e   ATAATAGAAG AG...TTGTT AAGAATGTTG TAAAAGATCC GACGGTAGAT
  M14   .......... .......... .......... .......... ..........
Prp2g   .......... .......... .......... .......... ..........

        451                                                500
 Paga   CAGGCG.... ........GT AGATTTAAGA GAGAAGATAA AGAAAATTAA
  73j   CAGGCGAAAA AATCGACGGA AGATTTAAAA GAGAGGTTAA AGGAAATTAA
Prp5e   CTGCCG.... ........GT AAATCTAACG CAGAAGTTAA AGAAAATTAA
  M14   .......... .......... .......... .......... ..........
Prp2g   .......... .......... .......... .......... ..........

        501                                                550
 Paga   AGAAGAATTA AATATAAGTG ACCCTTATTT TGATAAACAA TGGTATTTGG
  73j   AAAAGAATTA GGTATAAGTG ACCCTTGTTT TGATAAACAA TGGTATTTG.
Prp5e   AGAAGAATTA AATATAAGCA ACCCTTATTT TGATAAACAA TGGTATTTG.
  M14   .......... .......... .......... .......... ..........
Prp2g   .......... .......... .......... .......... ..........

        551                                                600
 Paga   TATAGTTTAT TCTTTTTTTC ATCAAAATTT GATTTTTTAA TTAGTTCAAT
  73j   .......... .......... .......... .......... ....TTTAAT
Prp5e   .......... .......... .......... .......... ....TTCAAT
  M14   .......... .......... .......... .......... ..........
Prp2g   .......... .......... .......... .......... ..........

        601                                                650
 Paga   AAGGATAAAG CTGGTGTAGA TATAAATGTT ACAGGTGTAT GGTTACAAGG
  73j   ACGGAAAAAC CTGGTGTAGA TATAAATGTT ACAGGTGTAT GGTTACAAG.
Prp5e   AAGGATAAAG CTGGTGTAGA TATAAATGTT ACAGGTGTAT GGTTACAAG.
  M14   .......... .......... .......... .......... ..........
Prp2g   .......... .......... .......... .......... ..........

        651                                                700
 Paga   TTTGATATTT GTGTTGTTAC TCGCCTTTTA ATGGATTTTA GGGATAAAGG
  73j   .......... .......... .......... .......... .GGATAACGG
Prp5e   .......... .......... .......... .......... .GGATAAAGG
  M14   .......... .......... .......... .......... ..........
Prp2g   .......... .......... .......... .......... ..........

        701                                                750
 Paga   GAAAAAATGT AACAGTTGCT ATTGTAGATG ATGGCTTAGA TTATACTAAC
  73j   GAAAAGGTGT AACAGTTGCC ATTGCAGATA ATGGCTTAGA TTATACTAAC
Prp5e   GAAAAAATGT AACAGTTGCT ATTGTAGATG ATGGCTTAGA TTATACTAAC
  M14   .......... .......... .......... .......... ..........
Prp2g   .......... .......... .......... .......... ..........

        751                                                800
 Paga   AAGGATTTGG CTCCAAATTA TGTTTGAAAA ACTATTATGG AAATCACTAT
  73j   AAGGATTTGG CTCCAAATTA T......... .......... ..........
Prp5e   AAGGATTTGG CTCCAAATTA T......... .......... ..........
  M14   .......... .......... .......... .......... ..........
Prp2g   .......... .......... .......... .......... ..........

        801                                                850
 Paga   TTTAACTTTT TTCAGAATGC TAACGCTTCA TATAATTTTG CTTCTAAAAC
  73j   .......... .....AATTC ACAGGGTTCA TATGATTTTG TTTCTAAAAC
Prp5e   .......... .....AATGC TAACGCTTCA TATAATTTTG CTTCTAAAAC
  M14   .......... .......... .......... .......... ..........
Prp2g   .......... .......... .......... .......... ..........
```

Figure 5

```
        851                                                900
  Paga  TGGCGACCCA AAACCTG... AACCTTCTGA CACGCATGGT ACTAAATGTG
   73j  TGACGACCCA AACCCTAAGA GCTCTTCTGA CACGCATGGT ACTAGATGTG
 Prp5e  TGGCGACCCA AAACCTG... GACCTTCGGA CACGCATGGT ACTAAATGTG
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        901                                                950
  Paga  CAGGAGAAGT GGCAGGCGCC AGGAATGATT TTTGTGGGCT TGGTGTTGCA
   73j  CAGGAGAAGT GGCAGCCGCC AGGAATGATT TTTGTGGGCT TGGTGTTGCA
 Prp5e  CAGGAGAAGT GGCAGGCGCC AGGAATGATT TTTGTGGGCT TGGTGTCGCA
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        951                                               1000
  Paga  TATGAATCTA ATATTTCAGG TATTTTTCTT TAATTGGTAC CTATCTAATA
   73j  TATGAATCTA ATATTTCAG. .......... .......... ..........
 Prp5e  TATGAATCTA ATATTTCAG. .......... .......... ..........
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        1001                                              1050
  Paga  TTGTTAAGGA TTACGATTTA TGCCTTCTGC TCGTTCGTCT TGGCTTGAAG
   73j  ........GA TTACGATTTT TGCCTTCTGG TCTCTCGTAT CATCTTGAGT
 Prp5e  ........GA TTACGATTTA TGCCTTCTGC TCGTTCGTCT TGGCTTGAAG
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        1051                                              1100
  Paga  GAGAAGCTCT TATTTACAAA TATGATGTTA ATCATATTTA TTCTTGTAGC
   73j  CACTAGCTCT TAGTTATAAA CCGAATGTTA ATTATATTTA TTCTTGTAGC
 Prp5e  GAGAAGCTCT TATTTACAAA TACGATGTTA ATCATATTTA TTCTTGTAGC
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        1101                                              1150
  Paga  TGGGGACCTG CCGATACTGG GAATTTAACT CAAGATATTT TTTATACTAC
   73j  TGGGGACCTC CTGGTGATGG ATATGCAGCT ATCCCAATGT ATCCTACTAC
 Prp5e  TGGGGACCCG CCGATACTGG GAATTTAACT CAAGATATTT TTTATACTAC
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        1151                                              1200
  Paga  TTATTCTGCA ATTATTAAAG GGATAAATCA AGGAAGGAAT GGTCTTGGTT
   73j  TTATTCTGCA ATTATTAAAG GGATAAAAGA AGGAAGGAAC GGTCTTGGCT
 Prp5e  TTATTCTGCA ATTATTGAAG GGATAAATCA AGGAAGGAAT GGTCTTGGTT
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        1201                                              1250
  Paga  CTATATACGT TTTCGGGTCA GGAAATGGTG GATATTTTGA TAATTGTAAT
   73j  CTATATATGT TTTTGGAACC GGAAATGGTG GATCATTGGA TGGTTGTAAT
 Prp5e  CTATATACGT TTTCGGGTCA GGAAATGGTG GATATTTTGA TAATTGTAAT
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........

        1251                                              1300
  Paga  TACGATGGAT ATGCAAATAG CCCATATACT ATTACTATCG CTGCTATAGA
   73j  TACGATGGAT ATGCAAATAG TCCATATACT ATTACTATCG CTGCTATAGA
 Prp5e  TACGATGGAT ATGCAAATAG CCCATATACT ATTACTATCG CTGCTATAGA
   M14  .......... .......... .......... .......... ..........
 Prp2g  .......... .......... .......... .......... ..........
```

Figure 5

```
       1301                                                   1350
 Paga  TGCAGAAGAA AAAAGATTCA TATTTTCAGA GCCATGTCCT TGTATTTTAG
  73j  TTCAGAAGAT AAAAATTTTT ATTTTTCAGA GTCATGTCCT TGCATTTTGG
Prp5e  TGCAGAAGAA AAAAGATTCA TATTTTCAGG GCCATGTCCT TGTATTTTAG
  M14  .......... .......... .......... .......... ..........
Prp2g  .......... .......... .......... .......... ..........

       1351                                                   1400
 Paga  CTTCTACGTA TTCTGGCAAG CGTGGTGCAT ATATTGTAAT CTTTTCTTTT
  73j  CTTCTACATA TTCTGGCGGA GAAAATGGAT CTATT..... ..........
Prp5e  CTTCTACGTA TTCTGGCAAG CGTGGTGCAT ATATT..... ..........
  M14  .......... .......... .......... .......... ..........
Prp2g  .......... .......... .......... .......... ..........

       1401                                                   1450
 Paga  TTTTTATAAT AAATTGATCG TTTTAGTATA CTACGGATGT TGGTACGACA
  73j  .......... .......... ......TATA CTACGGATCT TGGTAAGGAG
Prp5e  .......... .......... ......TATA CTACGGATGT TGGTACGACA
  M14  .......... .......... .......... .......... ..........
Prp2g  .......... .......... .......... .......... ..........

       1451                                                   1500
 Paga  GAATGCAGCA TTAGACATAC TGGAAGTTCT GCTTCTACAC CTCTTGCTGC
  73j  GGATGCACTA CTGAACATAC TGGAGCTTCT GCTTCTACAC CTCTTGCTGC
Prp5e  AAATGCAGCA TTAGACATAC TGGAAGTTCT GCTTCTACAC CTCTTGCTGC
  M14  .......... .......... .......... .......... ..........
Prp2g  .......... .......... .......... .......... ..........

       1501                                                   1550
 Paga  GGGTGTTATT GCTCTTCTTC TTTCAGCATG GTAAGAATAT CATTAAAATT
  73j  GGGTATTATT GCTCTTGTTC TTTCAGCGAA .......... ..........
Prp5e  GGGTGTTATT GCTCTTCTTC TTTCAGCATG .......... ..........
  M14  .......... .......... .......... .......... ..........
Prp2g  .......... .......... .......... .......... ..........

       1551                                                   1600
 Paga  ATTTGACTAA AAAATTAGTC CTAATCTTAC ATGGCGTGAT ATTCAAGCTT
  73j  .......... ........TC CTAATCTTAC ATGGCATGAT GTTCAAGCGT
Prp5e  .......... ........TC CTAATCTTAC ATGGCGTGAT ATTCAAGCCT
  M14  .......... .......... .......... .......... ..........
Prp2g  .......... .......... .......... .......... ..........

       1601                                                   1650
 Paga  TGATTGTGGA GACAGCTGTT CCATTTAATC CGAGTCATCC TGATTGGGAT
  73j  TGATTGTGGA AACAGCTGTT CCATTTAATT TGGAATATCC TGGATGGGAT
Prp5e  TGATTGTGGA GACAGCTGTT CCATTTAATC CGAGTCACCC TGATTGGGAT
  M14  .......... .......... .......... .......... ..........
Prp2g  .......... .......... .......... .......... ..........

       1651                                                   1700
 Paga  GATCTTCCTT CTGGACGTCG TTATAATAAT TTTTTCGGTT ATGGAAAACT
  73j  AAACTTCCTT CTGAACGTCA TTATAGTAAT AATTTTGGCT TTGGAAAGCT
Prp5e  GATCTTCCTT CTGGACGTCG TTATAATAAT TTTTTCGGTT ATGGAAAACT
  M14  .......... .......... .......... .......... ..........
Prp2g  .......... .......... .......... .......... ..........

       1701                                                   1750
 Paga  AGATGCATAT AGAATGGTCG AAAAAGCAAG AACATTTAAA ACCTTAAATC
  73j  AGATGCGTAT AGAATGGTCG AAAGAGCAAA AACATTTAAA ACATTAAATG
Prp5e  AGATGCATAT AGAATGGTCG AAAAAGCAAG AACATTTAAA ACCTTAAATC
  M14  .....CATAT AGAATGGTCG AAAGAGCAAA AACATTTAAA ACATTAAATG
Prp2g  .......... .......... .......... .......... ..........

       1751                                                   1800
```

17/21

Figure 5

```
 Paga   CTCAGACAAT GTTTTCAACT CAACTAATAC CACTTAATAA GAAATTTTCT
  73j   CTCAGACAAT GTTTTCAACT CAACTAATAC CACTTAATAA GACATTTTCT
Prp5e   CTCAGACAAT GTTTTCAACT CAACTAATAC CACTTAATAA GAAATTTTCT
  M14   CTCAGACAAT GTTTTCAACT CAACTAATAC AAATTAATAT GAAATTTCCT
Prp2g   .......... .......... .......... .......... ..........

        1801                                             1850
 Paga   GAAAACGGTG GGCATATCAC AAGCAGTTTT TATATTCATC GTGGATATCC
  73j   GAAAACGGTG GGCATATCAC AAGCACTTTT TATATTGATA GTGGATCTCC
Prp5e   GAGAACGGTG GGCATATCAC AAGCAGTTTT TATATTCATC GCGGATATCC
  M14   GATCCCAGTA GACGTATCAC GAGCAGTTTT TATATTCATA GTGGATATCC
Prp2g   .......... .......... .......... .......... ..........

        1851                                             1900
 Paga   TAAGCATTAT AAATTTAAAA GTTTAGAGTA TGTTGGTGTT TCATTTCATT
  73j   TACGCATTAT AACTTTAAAA ATTTGGAATA TGTTGGTGTT TCATTTCATT
Prp5e   TAAGCATTA. .......... .......... .......... ..........
  M14   TACGCATTAT AACTTTAAAA ATTTGGAATG TGTTGGTGTT TCATTTCATT
Prp2g   .......... .......... .......... .......... ..........

        1901                                             1950
 Paga   ATCAGCACCA AAGAAGAGGT CATCTAGAGT TTAATATTAC CAGTCCTTCT
  73j   ATAAGCACCA ATATAAAGGT CATCTGGAGT TTAATATTAC CAGTCCTTCT
Prp5e   .......... .......... .......... .......... ..........
  M14   ATCAGCACCA AAAAAGAGGT CGTCTGGAGT TTAGTATTAC AAGCCCTGCT
Prp2g   .......... .......... .......... .......... ..........

        1951                                             2000
 Paga   GGAGTTACTT CAGTATTAGC ACATAGACGT AATCGTGATA AACATGGTGG
  73j   GGAGTTACTT CAGTATTAGC ACATAGACGT ATTAATGATT ATAATAGTGG
Prp5e   .......... .......... .......... .......... ..........
  M14   AATGTTACTT CAAAATTAGC ACGTGTACGT GTTCGTGATG AAGAAAGTGG
Prp2g   .......... .......... .......... .......... ..........

        2001                                             2050
 Paga   CAGTATTCTT TGGACTTTTA TGACTGTAAA GCATTGGTAT TTTGTTTCAT
  73j   CACTTTTCAT TGGTTTTTTA CGACTGTAAA GCATTG.... ..........
Prp5e   .......... .......... .......... .......... ..........
  M14   CACTTTTTCT TGGATTTTTA CGACTGTAAA GCATTG.... ..........
Prp2g   .......... .......... .......... .......... ..........

        2051                                             2100
 Paga   TTTGTAAAAT AATAACTAAT GATTTTAGGG GAGAATCCAT TGTAGGTAAT
  73j   .......... .......... ........GG GAGAAACCAT TGTAGGTAAC
Prp5e   .......... .......... .......... .......... ..........
  M14   .......... .......... ........GG GGGAAAAGAT TGTAGGTAAT
Prp2g   .......... .......... .......... .......... ..........

        2101                                             2150
 Paga   TGGACTATCG ATGTTGAAGA TAAAAAGGAT GAGAATCTAG ATGGTGGAGT
  73j   TGGACTATCG ATGTTGAAGA TGAAAAGGTT TCGAATCTAG ATGGTGAAAT
Prp5e   .......... .......... .......... .......... ..........
  M14   TGGACTATCG ATGTTGAAGA TGAAAAAGAT CCGAATCTAG ATGGTGAAGT
Prp2g   .......... .......... .......... .......... ..........

        2151                                             2200
 Paga   TTTTGATTGG CAACTTCATT TTTTCGGGGA GTCTTGTGAA TCA...GAAG
  73j   TTTTGATTGG CAACTTCATT TTTTCGGAGA GTCTATTGAT TCAAGTAAAG
Prp5e   .......... .......... .......... .......... ..........
  M14   TTTTAATTGG CAACTTCATT TTTTCGGAGA GTCTATTGAT TCAACAAAAG
Prp2g   .......... .......... .......... .......... ..........

        2201                                             2250
 Paga   GCGTACCGCC TCCTTCATAT CCTTTTCTAT CTAGATATCC AACTACTACG
```

18/21

Figure 5

```
  73j   CAGAACTTCA TCCTCCATAT CCTTTTAAGC CTCAA..... ..........
Prp5e   .......... .......... .......... .......... ..........
  M14   CACA...GCC TCCTCCATAT CCTTTTGTGC ATAAACAACC AACTACTATG
Prp2g   .......... .......... .......... .......... ..........

        2251                                            2300
 Paga   CCTCCACCAG ATCCAGATGC TACACCTTCT CCAGATCTGG ATGCTAACCT
  73j   .......... .......... .......... .......... ..........
Prp5e   .......... .......... .......... .......... ..........
  M14   CCTCCGCCAG AACCAACTAC TACGCTTCCA TCAGATCCAG ATGCTACATC
Prp2g   .......... .......... .......... .......... ..........

        2301                                            2350
 Paga   TCAGCCAGAT TCAAATGCTG ACTCT..... .......... .........C
  73j   .......... .......... .......... .......... ..........
Prp5e   .......... .......... .......... .......... ..........
  M14   TCTACCAGAT TTAAATGTTG CACCTTCGCC AGATTTAAAT GCTAACCCTC
Prp2g   .......... .......... .......... .......... ..........

        2351                                            2400
 Paga   AACCTCAACC AGATGTTAAG CCTCTGCCTT CATTAGATAT TGAGCCTCAA
  73j   .......... .......... .......... .......... ..........
Prp5e   .......... .......... .......... .......... ..........
  M14   AACCTCAACC AGATCCTGGG TCTCCGCCCT CATCAGATCC TGAGTCTCCG
Prp2g   .......... .......... .......... .......... ..........

        2401                                            2450
 Paga   CCTCCATCAG AACCAGATTC TAACCCTCCA TCAGATCTTA GCTCTCAGCA
  73j   CCTCCTTCAA AACCTGCGCC TCCATCAAAA CCAGATCCTA ACCCTCCATC
Prp5e   .......... .......... .......... .......... ..........
  M14   TCTTCATTAG AACCTGCGCC TCCATCAAAA CCAGATCCTA ACCCTCCATC
Prp2g   .......... .......... .......... .......... ..........

        2451                                            2500
 Paga   AGATCC.... .......... ....AGATAC TTCGCTTTCA TCAAATGCAA
  73j   AGATCCTAGC TCTCAGCAAG ATTCAGATAC TTCGCTTTCA TCAACTCCAA
Prp5e   .......... .......... .......... .......... ..........
  M14   AGATCCTAGC TCTCAGCAAG ATCCAGATAC TTCGCTTTCA TCAAATCCAA
Prp2g   .......... .......... .......... .......... ..........

        2501                                            2550
 Paga   CTTCTACATC TTCATCAGAA CTACCACCAC TACCACCACC ACCGCCGCCA
  73j   CTTCTACATC TTCATCAAAA .......... .......... ..........
Prp5e   .......... .......... .......... .......... ..........
  M14   CTTCTACATC TTCATCAGAA CCACCACCAC TACCACCACC ACCGCCAC..
Prp2g   .......... .......... .......... .......... ..........

        2551                                            2600
 Paga   CCTGCACCTG CACCACCTGC ACCTGCACCA CCTCCACCAC CGCCGCCACC
  73j   .......... .......... .......... .......... ..........
Prp5e   .......... .......... .......... .......... ..........
  M14   .CTGCACCTG CACCGCCTCC ACCACCGCCG CCACCACCAT CTCGGCCGGA
Prp2g   .......... .......... .......... .......... ..........

        2601                                            2650
 Paga   ACCACCTCGG CCGGAACCAC AACCACAACC AGAGACACAA CCAGAGACAC
  73j   .......... .......... .......... .......... ..........
Prp5e   .......... .......... .......... .......... ..........
  M14   ACCAGAACCA GAACCGCGAC CAGAACCAAA ACCAAAACCA GAACCAGAAC
Prp2g   .......... .......... .......... .......... ..........

        2651                                            2700
 Paga   AACCAGAGAC ACAACCAGAG ACACAACCAG AGACACAACC ACCACAACCA
  73j   .......... .......... .......... .......... ..........
```

Figure 5

```
 Prp5e  .......... .......... .......... .......... ..........
   M14  CAGAACCAGA ACCAGAACCA GAACTAGAAC TAGAACTAGA ACTAGAACTA
 Prp2g  .......... .......... .......... .......... ..........

        2701                                             2750
  Paga  CCACAACCAC CACAATCAGA GACACAACCA GAACCAGAAC CAGAACCAGA
   73j  .......... .......... .......... .......... ..........
 Prp5e  .......... .......... .......... .......... ..........
   M14  GAACCAGAAC CAGAACCAGA ACCAGAACCA GAACCAGAAC CAGAACCAGA
 Prp2g  .......... .......... .......... .......... ..........

        2751                                             2800
  Paga  ACCAGAACCA GAACCAGAGC CAGAGCCAGA GCCACAACCA GAACCAGAAC
   73j  .......... .......... .......... .......... ..........
 Prp5e  .......... .......... .......... .......... ..........
   M14  GCCACAACCA GAGCCACAAC CAGAGCCACA ACCACAACCA GAGCCACAAC
 Prp2g  .......... .......... .......... .......... ..........

        2801                                             2850
  Paga  CAGAGACACA ACCAGAGCCA CAACCACCAC AACCAGAGCC ACAACCACCA
   73j  .......... .......... .........C TGTCACCACC ACCTACACCT
 Prp5e  .......... .......... .......... .......... ..........
   M14  CAGAGCCACA ACCACAACCA GAGCCACAAC CAGAGCCACA ACCACAACCA
 Prp2g  .......... .......... .......... .......... ..........

        2851                                             2900
  Paga  CAACCAGAGC CACAACCAGA GCCACCTGCA TCTCCACCAA AACTACAACC
   73j  CAACCAAAGC CAGAACCACA ACCGGAACAG AAACCGACAT CAATAGCTTC
 Prp5e  .......... .......... .......... .......... ..........
   M14  CCGCTGCCAC AACCACCGCT GCCACCTGCA CCTCCACCAA AACCACAACC
 Prp2g  .......... .......... .......... .......... ..........

        2901                                             2950
  Paga  GGAACAAAAA CCAACATCAA TAACTTCATC TACATCTACG ACTTCATCGA
   73j  ATCTACAACA TCAACTAATT TAATTCCACC AGCTCCCACA TCTTCATCAA
 Prp5e  .......... .......... .......... .......... ..........
   M14  GGAACAAAAA CCAACATCAA TAACTTCATC TACATCTACG ACTTCATCGA
 Prp2g  .......... .......... .......... .......... .....ATCAA

        2951                                             3000
  Paga  GCAAAACTAA AATATCAACC ACTCGAAAAG CTTCATGTAC TAT.......
   73j  GCAAAACTAA AACATCAACC ACTCGAAAAG CTTCATCTAC TA........
 Prp5e  .......... .......... .......... .......... ..........
   M14  GCAAAACTAA AATATCAACC ACT....... .......... ..........
 Prp2g  GCAAAACTAA AATATCAACC ACTCGAAAAG CTTCATCTAC TAAAACTTCA

        3001                                             3050
  Paga  ........AA CAGTCTTTAT AGGGCCATCT CCTACTGAGG GTGTTTCTAC
   73j  .......CAA AAACCTCTAC ACGGCCGTCT CCTACTGAGG GTACTTTTAC
 Prp5e  .......... .......... .......... .......... ..........
   M14  .......... .......... .......... .......... ..........
 Prp2g  TCTACTACAA AAACTTCTGC ACGGCCGTCT CCTACTGAGG GTACTTTTAC

        3051                                             3100
  Paga  TGGATCAAGT GCTTCTCATC TTTCATTCTT CGAAAAAAGG CATTTGTTAC
   73j  TGGATCAGGC TGTTCTCATC TTTCATTCTT CGAAAAAAGG CATTTGTTAC
 Prp5e  .......... .......... .......... .......... ..........
   M14  .......... .......... .......... .......... ..........
 Prp2g  TGGATCAAGT GCTTCTCGTC TTTCATTCTT CGAAAAAAGG CATTTGTTAC

        3101                                             3150
  Paga  TTCAAATGAT ATTATTGTTA TTCTTTTTCT TATTTTTGGG TTACTCTTTT
   73j  TTCAGATGAT ATTATTGTTA TTCTTTTTCT TATTTTTGGG TTACTCTTTT
 Prp5e  .......... .......... .......... .......... ..........
```

Figure 6

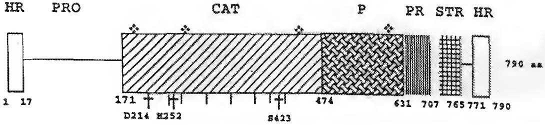
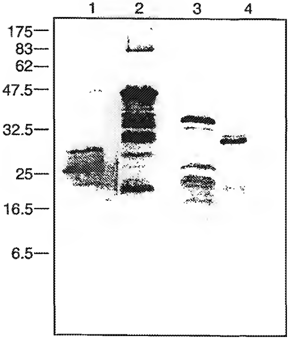

HR
PRO
CAT
P
PR
STR
HR
790 aa
1
17
171
D214
H252
S423
474
631
707
765
771
790

Fig.7.



INTERNATIONAL SEARCH REPORT

national Application No
PCT/GB 98/00704

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12N9/58 C12N15/55 C07K16/14

According to International Patent Classification(IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12N C07K

Documentation searched other than minimumdocumentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WADA M ET AL: "MSG gene cluster encoding major cell surface glycoproteins of rat Pneumocystis carinii" DNA RESEARCH, vol. 1, no. 4, 1994, TOKYO JP, pages 163-168, XP002071766 cited in the application see the whole document	1-7,22
A	MASSETTI A P ET AL: "Identification of Pneumocystis carinii proteases with a role in adhesion mechanisms" IXTH INTERNATIONAL CONFERENCE ON AIDS, vol. 0, no. 0, 6 - 11 June 1993, BERLIN DE, page 388 XP002071767 see abstract nr.: PO-B10-1515	1

-/--

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publicationdate of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of theinternational search 23 July 1998	Date of mailing of the international search report 05/08/1998
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Authorized officer De Kok, A

1

INTERNATIONAL SEARCH REPORT

national Application No
PCT/GB 98/00704

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	DATABASE WPI Section Ch, Week 9105 Derwent Publications Ltd., London, GB; Class B04, AN 91-033527 XP002071770 & JP 02 303 498 A (NIPPON KAYAKU KK) see abstract	10-12,21
A	WO 96 30004 A (UNIVIVERSITY OF CALIFORNIA) 3 October 1996 see page 4, line 20 - page 5, line 5	13-20
A	WO 91 02092 A (GENE TRAK SYSTEMS) 21 February 1991 see page 1 - page 7	1,2
A	WO 93 07274 A (THE GENERAL HOSPITAL CORP) 15 April 1993 see the whole document	1-21
P,X	WADA M ET AL: "cDNA cloning and overexpression of cell surface subtilisin-like proteases (SSP) of Pneumocystis carinii" THE JOURNAL OF EUKARYOTIC MICROBIOLOGY, vol. 44, no. 6, November 1997, US, pages 54S-56S, XP002071768 see abstract	1-7,17, 22
P,X	LUGLI E B ET AL: "A Pneumocystis carinii multi-gene family with homology to subtilisin-like serine proteases" MICROBIOLOGY, vol. 143, no. 7, July 1997, READING GB, pages 2223-2236, XP002071769 cited in the application see the whole document	1-7,22

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/GB 98/00704

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9630004 A	03-10-1996	US 5739170 A	14-04-1998
		CA 2215245 A	03-10-1996
		EP 0817624 A	14-01-1998
WO 9102092 A	21-02-1991	AT 121793 T	15-05-1995
		AU 6356390 A	11-03-1991
		CA 2035872 A	12-02-1991
		DE 69018961 D	01-06-1995
		DE 69018961 T	23-11-1995
		EP 0438587 A	31-07-1991
		JP 4501211 T	05-03-1992
		US 5519127 A	21-05-1996
WO 9307274 A	15-04-1993	US 5442050 A	15-08-1995
		AU 2869192 A	03-05-1993